Strategic Disinformation Generation and Detection

Wenxiao Yang[1], Yunfei (Jesse) Yao[2], and Pengxiang Zhou[*3]

[1]University of California, Berkeley, ywenxiao@berkeley.edu
[2]The Chinese University of Hong Kong, jesseyao@cuhk.edu.hk
[3]University of Southern California, pzhou584@usc.edu

July 20, 2025

Abstract

Disinformation detection is becoming increasingly important and relevant because it is easier than ever to create and disseminate disinformation. How does detection ability affect the incentive to generate disinformation? Given the practical constraints of classification technology, how should a detector be designed? To answer these questions, this paper studies the problem where a sender strategically communicates his type (high or low) to a receiver, and a lie detector generates a noisy signal on the truthfulness of the sender's message. The receiver then infers the sender's type both through the message from the sender and through the signal from the detector. We find a non-monotonic relationship between the probability that the low-type sender is lying and the accuracy of detection. More accurate detection (a higher true-positive rate and a lower false-positive rate) increases the probability of lying when the true-positive rate is low, because of a persuasive effect. By contrast, more accurate detection decreases the probability of lying when the true-positive rate is high, because of a dissuasive effect. We also characterize the optimal detector design. The designer always chooses the lowest feasible false-positive rate for any true-positive rate. The possibility of false-positive alarms implies that the designer chooses an intermediate true-positive rate rather than the highest true-positive rate. Counter-intuitively, the optimal detector may raise an alarm about a smaller percentage of disinformation when its underlying classifier is better at distinguishing the sender's type.

[*]Corresponding author. Authors are ordered alphabetically. We thank Jieteng Chen, Jiajia Cong, Liang Guo, Ganesh Iyer, Yuichiro Kamada, Sam Kapon, Song Lin, Ram Rao, Zuo-Jun (Max) Shen, J. Miguel Villas-Boas, Keyan Zhu, and Zihao Zhou for helpful comments and suggestions. We also thank seminar participants at NYU Shanghai and the Summer Institute in Competitive Strategy 2025.

"It is better that ten guilty persons escape than that one innocent suffer."

— William Blackstone, *Commentaries on the Laws of England*

1 Introduction

There is widespread disinformation nowadays, including fake reviews, ad fraud, manipulated transactions, fraudulent resumes, and misleading posts (Anderson and Simester, 2014; Mayzlin, Dover, and Chevalier, 2014; Luca and Zervas, 2016; Gordon et al., 2021; He et al., 2022; He, Hollenbeck, and Proserpio, 2022).[1] The emerging generative AI technologies further exacerbate such deceptive practices. These issues have attracted much attention in places such as online platforms, political realms, and justice systems, where trust and integrity are paramount. Yet, detecting disinformation remains challenging (Callander and Wilkie, 2007; Dziuda and Salas, 2018; Mattes, Popova, and Evans, 2023). In response to the ubiquitous deceptive activities, platforms and regulators have devised various ways of detecting and raising an alarm about disinformation, usually with the help of sophisticated algorithms. For example, Yelp utilizes automated systems to identify compensated or incentivized reviews and flags businesses with suspicious activities.[2] Using an internal system, Twitter labels false or misleading content to help people "find credible and authentic information" and "make informed decisions."[3] To fight against fake accounts and fraudulent activities, LinkedIn has built "automated detection systems at scale."[4] Such detection and alarm attempts help individuals decide whether they will go to a particular restaurant, re-post a social media post, or connect with a Linkedin account.

When developing mechanisms to detect deceptive information, the designer typically faces a dilemma between increasing the likelihood of correctly recognizing deceptive content (true positives) and reducing the probability of falsely identifying genuine content as deceptive (false positives). Such a trade-off between Type I error (false-positive) and Type II error (false-negative) is a well-known statistical challenge faced by many fields (Goodin, 1985; Buckland and Gey, 1994; Lieberman and Cunningham, 2009; Cappelen, Cappe-

[1] There are two related terms commonly used in the media and the literature - misinformation and disinformation. According to Dictionary.com, misinformation refers to "false information that is spread, regardless of whether there is intent to mislead", whereas disinformation means "deliberately misleading or biased information." By focusing on the strategic incentive of the information provider (sender), this paper studies disinformation. We thank one of the anonymous reviewers for raising this important distinction between misinformation and disinformation.

[2] https://trust.yelp.com/trust-and-safety-report/2023-report/ and https://trust.yelp.com/consumer-alerts/quarterly-alerts/.

[3] https://blog.x.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.

[4] https://www.linkedin.com/blog/engineering/trust-and-safety/automated-fake-account-detection-at-linkedin.

len, and Tungodden, 2023). In a recent podcast interview, Mark Zuckerberg acknowledged that Facebook's content detection and moderation team is well aware of the precision-recall trade-off and actively seeks to balance correctly recognizing deceptive content with reducing erroneous actions on genuine content.[5]

How does the detection ability affect the incentive to generate disinformation? Given the practical constraints of classification technology, how should the detectors be designed? To answer these questions, this paper studies the problem where a sender strategically communicates his type (high or low) to a receiver, and a lie detector generates a noisy signal on the truthfulness of the sender's message.[6] The receiver then infers the sender's type both through messages from the sender and through signals from the detector. Previous work has considered the possibility that a detector may fail to send an alarm when there is disinformation (false negative). Observing that the detector may make another type of mistake by sending a false alarm in the absence of disinformation (false positive), a key contribution of our paper is to allow for both types of mistakes in disinformation detection. In addition to being more realistic, it also leads to qualitatively different insights about the relationship between the probability that the low-type sender is lying and the accuracy of detection. The other main contribution of this paper is to endogenize the design of the detector rather than treating the detection technology as exogenously given. The optimal detector design is also qualitatively different with and without consideration of false-positive alarms.

Specifically, this paper considers a model where a receiver makes a binary decision between actions r_H and r_L. The sender may be either the H type (the high type) or the L type (the low type); this is his private information. The sender always wants the receiver to take action r_H, whereas the receiver prefers to take action r_H if the sender's type is H and to take action r_L if the sender's type is L. The sender can send a strategic message about his type (m_H for high type and m_L for low type) to the receiver, while a lie detector generates a noisy signal on the truthfulness of the sender's message. The receiver infers the sender's type both through the message from the sender and through the signal from the detector, and then makes a decision.

Because of practical limitations, the detector may make two types of mistakes. It may fail to send an alarm when the sender lies (false negative). It may also send a false alarm when the sender is truthful (false positive). (CMA, 2015; Lappas, Sabnis, and Valkanas, 2016). Previous work has focused on the first type of mistake by implicitly assuming that the false-positive rate is zero (Becker and Stigler, 1974;

[5]Recall is one minus the Type II error rate, whereas precision is inversely related to the Type I error rate. The link to the podcast interview is https://www.youtube.com/watch?v=7k1ehaE0bdU.

[6]We refer to the sender as "he" and the receiver as "she" throughout the paper

Dziuda and Salas, 2018; Balbuzanov, 2019). Given the practical constraints of classification technology, however, the sender cannot avoid making the second type of mistake (false positive) unless he never sends an alarm.[7] Moreover, false positives are not only ubiquitous but also economically significant. J.P. Morgan views false positives as a multi-billion dollar problem.[8] Global business loses more than $100 billion every year because of false positives, which is even more than the actual fraud costs.[9] In order to understand the strategic impact of disinformation detection in a realistic setting, this paper explicitly considers the possibility of false-positive alarms.

We first study how the detection technology affects the equilibrium outcomes. We find a non-monotonic relationship between the probability that the low-type sender is lying and the accuracy of detection: more accurate detection (a higher true-positive rate and a lower false-positive rate) increases the probability of lying when the true-positive rate is low and decreases it when the true-positive rate is high. Two effects drive the non-monotonicity. Because the detector is more likely to send no alarm when the sender is high-type than when he is low-type, the receiver becomes more certain that the sender is high-type if she receives no alarm. The presence of a detector persuades the receiver to trust the sender's m_H message more in this case. We call this posterior belief-enhancing effect a *persuasive effect*. Because the detector is more likely to send an alarm when the sender is low-type than when he is high-type, the receiver becomes more certain that the sender is low-type if she receives an alarm. The presence of an alarm causes the receiver to have less trust about the sender's m_H message. We call this posterior belief-reducing effect an *dissuasive effect*. When the true-positive rate is low, the receiver adopts a mixed strategy between actions r_H and r_L after observing message m_H and no alarm.[10] As the detector becomes more accurate, for a fixed sender's strategy, the receiver's posterior belief after observing no alarm will be higher because of the larger persuasive effect. So, the low-type sender can afford to lie more frequently in equilibrium. When the true-positive rate is high, the detector will catch a high proportion of low-type senders who are lying. This creates a low incentive for lying. Consequently, the receiver always takes the sender's desired action if there is no alarm and adopts a mixed strategy after seeing an alarm. As the detector becomes more accurate, for a given sender's strategy, the receiver's posterior belief after observing an alarm will be lower because of a larger dissuasive effect. In equilibrium, the low-type sender needs to lie less frequently in order for there to be a positive probability

[7] Similarly, the sender cannot avoid making the first type of mistake (false negative) unless he always sends an alarm.

[8] https://www.jpmorgan.com/insights/payments/analytics-and-insights/cnp-fraud-prevention-combat-chargebacks.

[9] https://www.vesta.io/blog/false-positives-and-how-to-prevent-them.

[10] In other words, the receiver takes action r_H with some probability and takes action r_L with some probability after observing message m_H and no alarm.

that the receiver will take the sender's desired action even after observing an alarm.

We then characterize the optimal detector design. The receiver and both types of sender all benefit from a lower false-positive rate, whereas the low-type sender is hurt by a higher true-positive rate. Therefore, the designer always chooses the lowest feasible false-positive rate for any given true-positive rate. The possibility of false-positive alarms implies that the designer will not choose the largest true-positive rate. Instead, the designer chooses different intermediate true-positive rates for different objectives. Counter-intuitively, the optimal detector may raise an alarm about a smaller percentage of disinformation when its underlying classifier is better at distinguishing the sender's type.

As the first attempt to incorporate false-positive errors in disinformation detection, this paper seeks to offer a general understanding of their economic implications. We examine various objective functions of the detector designer, including maximizing the receiver's payoff, the sender's payoff, and a weighted average of the players' payoffs (including social welfare). While these objectives are relevant to some practical applications, we acknowledge that real-world designers may act more strategically. The implications of such strategic behavior depend on the institutional details of the problems. In an extension, we analyze a particular setting in which an e-commerce platform (designer) chooses the price (i.e., commission fee) in addition to the information structure of the detector. Pricing and detector design jointly affect the seller's (sender's) platform entry decision, which in turn affects the designer's payoff. Joint consideration of pricing and information communication offers richer managerial insights and demonstrates how our main model can be adapted to accommodate more strategic objectives, contingent on the specific institutional context.

Related Literature

Our research is most closely related to the strategic communication literature. One stream of the literature on verifiable disclosure, initiated by Grossman (1981) and Milgrom (1981), assumes that information is verifiable and thus agents can withhold it but cannot lie. Another stream of the literature on cheap talk, developed by Crawford and Sobel (1982), considers a model where information is unverifiable and thus agents can freely send deceptive messages. Later work studies strategic communication in markets where firms and consumers interact (Villas-Boas, 2004; Shin, 2005; Guo, 2009; Guo and Zhao, 2009; Kuksov, 2009; Kuksov and Lin, 2010; Mayzlin and Shin, 2011; Sun, 2011; Zhang, 2013; Branco, Sun, and Villas-Boas, 2016; Iyer and Singh, 2018; Sun and Tyagi, 2020; Wu, Zhang, and Xie, 2020; Iyer and Singh, 2022; Lauga, Ofek, and Katona, 2022; Zheng and Singh, 2023; Chen, Huang, and Gong, 2024; Lee, Shin, and Yu, 2024; Qiu

and Rao, 2024; Chen, Du, and Lei, 2025; Ning, Shin, and Yu, 2025). In the verifiable disclosure literature, senders can be viewed as having an infinite lying cost and therefore never lie, whereas they have zero lying cost in the cheap talk literature. The more recent literature on the theory of costly lying (Kartik, Ottaviani, and Squintani, 2007; Kartik, 2009) assumes that the sender has a finite lying cost and can be viewed as the middle ground of two extreme cases. The presence of lying cost in the above papers allows messages to have a signaling role. However, the information is still completely unverifiable. Recent work (Dziuda and Salas, 2018; Balbuzanov, 2019) starts considering the possibility of an imperfect detector that may detect the lie with some probability. In such cases, the sender's message becomes partially verifiable.

Because of practical limitations, the detector may make two types of mistakes. It may fail to send an alarm when the sender lies (false negative). It also may send a false alarm when the sender is telling the truth (false positive). By assuming that the detector detects the lie with some probability, previous work implicitly assumes that there is no false positive (the false-positive rate is zero). A major contribution of our paper is to allow for both types of mistakes by studying a detector with general true-positive and false-positive rates. Type I and type II errors may be viewed conceptually as similar to one another in that both types of errors make information less precise. Nevertheless, this paper shows that they generate different effects on strategic communication, which in turn provides important implications for designing the disinformation detector. The other key contribution is to endogenize the design of the detector rather than treating the detection technology as exogenously given. One reason that previous literature has focused on exogenous detectors is that, in the absence of false positives, the receiver's payoff, the high-type sender's payoff, and social welfare are all (weakly) increasing in the true-positive rate of the detector; we will discuss this as a benchmark situation in section 3.2. So, there is no trade-off, and the designer always wants to maximize the true-positive rate of the detector. In contrast, we will show that, in the presence of false-negative alarms, the designer prefers an intermediate true-positive rate to the highest true-positive rate. In this case, the optimal detector design becomes both non-trivial and managerially important.

We use an information design framework to study the general design of the detector. Since Rayo and Segal (2010) and Kamenica and Gentzkow (2011) initiated the study of the optimal design of flexible information provision with commitment, researchers have found its applications in various areas, including advertising, recommendation algorithms, influencer marketing, search, and online platforms (Jerath and Ren, 2021; Berman, Zhao, and Zhu, 2022; Du and Lei, 2022; Iyer and Zhong, 2022; Ke, Lin, and Lu, 2022; Pei and Mayzlin, 2022; Shin and Wang, 2024; Shulman and Gu, 2024; Yao, 2024). Unlike papers in this

literature, the information design problem in our model is just a subgame of a costly signaling game. Because of the presence of the sender's private information, the sender sends strategic signaling messages to influence the receiver's decision, on top of the information design of the detection technology. In addition, we are studying the design of detection technology rather than the information itself.

Our paper is also related to the growing literature on strategic interactions between humans and algorithms. Liang (2019); Miklós-Thal and Tucker (2019); Calvano et al. (2020); Salant and Cherry (2020); O'Connor and Wilson (2021) and Montiel Olea et al. (2022) study competitive dynamics among multiple algorithms. Berman and Katona (2013, 2020) study the impact of online algorithms on advertisers' and social media users' behavior. Lin, Shi, and Sun (2025) study the impact of Generative AI on consumers' search and purchasing decisions in online shopping platforms. Eliaz and Spiegler (2019) and Björkegren, Blumenstock, and Knight (2020) look at algorithm design with information manipulation by strategic agents. Qian and Jain (2024) investigate the impact of recommendation systems on digital content creation. Iyer and Ke (2024) study on strategic model selection in competitive environments. Iyer, Yao, and Zhong (2024) examine the precision-recall trade-off in the deployment of machine learning algorithms for targeting. A recent paper by Chen, Ke, and Shin (2025) studies the interaction between content creators' incentives to adopt AI and the platform's detection strategy that distinguishes AI-generated from human-created content. We focus instead on the design of a disinformation detector in a strategic communication game.

Lastly, our paper is related to the literature on information misrepresentation. Anderson and Simester (2014); Mayzlin, Dover, and Chevalier (2014); Lappas, Sabnis, and Valkanas (2016) and Luca and Zervas (2016) provide empirical evidence about the prevalence of strategic review manipulation. Mayzlin (2006) and Dellarocas (2006) theoretically study firms' costly misrepresentation of product quality. A growing literature investigates deceptive advertising practices that promote false claims about product quality (Piccolo, Tedeschi, and Ursino, 2015; Zinman and Zitzewitz, 2016; Rao and Wang, 2017; Piccolo, Tedeschi, and Ursino, 2018; Rhodes and Wilson, 2018). Jin, Yang, and Hosanagar (2023) examines a widespread phenomenon on e-commerce platforms where sellers place fake orders to boost the search ranking of their products. Some papers also examine the regulation and policy implications of deceptive activities (Piccolo, Tedeschi, and Ursino, 2015; Rhodes and Wilson, 2018; Papanastasiou, 2020; Chen and Papanastasiou, 2021). We contribute to this literature by considering the trade-offs between the false-positive and true-positive rates of the detection algorithm and studying the interaction between the sender's strategic communication strategy and the detection technology.

The rest of this paper is organized as follows. Section 2 introduces the main model. Section 3 presents several benchmarks. Section 4 solves the equilibrium and compares the results with the benchmarks. Section 5 studies two extensions. Section 6 concludes.

2 Model

2.1 States, Actions, and Payoffs

There is a sender (S), a receiver (R), and a designer. The receiver makes a binary decision between actions r_H and r_L. Depending on the specific applications, the receiver's action can be purchasing a product from an e-commerce seller, visiting a restaurant, re-posting social media content, clicking on an email link, sending a business contact request, etc. The sender is the H type with probability ρ and the L type with probability $1-\rho$ (we will use type H/L and high-type/low-type interchangeably throughout the paper). The sender's type is his private information and can be interpreted as the quality of an online marketplace seller's product, the quality of a restaurant, the trustworthiness of a social media content creator, the credibility of an email sender, the authenticity of an online business account, etc. The sender always wants the receiver to take action r_H, whereas the receiver prefers to take action r_H if the sender's type is H and to take action r_L if the sender's type is L. Table 1 summarizes players' payoffs.

(sender payoff, receiver payoff)	action r_H	action r_L
type H sender	$(\Delta_H^S > 0, \Delta_H^R > 0)$	$(0,0)$
type L sender	$(\Delta_L^S > 0, -\Delta_L^R < 0)$	$(0,0)$

Table 1: Players' Payoffs

The sender always earns a positive payoff if the receiver takes action r_H, $\Delta_H^S, \Delta_L^S > 0$. The receiver gains a positive payoff Δ_H^R if she takes action r_H when the sender's type is H, and suffers from a utility loss of Δ_L^R if she takes action r_H when the state is L. Both players' payoffs are normalized to zero if the receiver takes action r_L. Denote by $\hat{\rho}$ the *critical belief* such that the receiver is indifferent between taking either action, $\hat{\rho}\Delta_H^R - (1-\hat{\rho})\Delta_L^R = 0$. We study the non-trivial case where the receiver will take action r_L without any information by assuming that $\rho\Delta_H^R - (1-\rho)\Delta_L^R < 0 \Leftrightarrow \rho < \hat{\rho}$. We also focus on the case where $\Delta_L^S - \Delta_L^R \leq 0$, which means that successful lying by a low-type sender does not increase social welfare.

The sender can send a non-verifiable but detectable message $m \in \{m_H, m_L\}$ about his type to the

receiver. The sender is lying if his message is not aligned with his type (i.e., sending message m_H if his type is L or sending message m_L if his type is H). Consistent with previous literature, the sender needs to incur a positive cost of C if he lies.[11] The lying cost may come from the sender's intrinsic aversion to lying (Gneezy, 2005), the potential ex-post penalty for lying, or the effort of manipulating the information. We rule out the uninteresting case where the sender never lies because of a high lying cost by assuming that $C < \min\{\Delta_H^S, \Delta_L^S\}$.

A lie detector generates a noisy signal $l \in \{a, na\}$ on the truthfulness of the sender's message if the sender sends m_H. The detector will send an alarm, $l = a$, to the receiver if it thinks the message m_H is sent by a type L sender. It will send a no-alarm signal, $l = na$, to the receiver if it thinks the message m_H is sent by a type H sender or the message is m_L.

Eventually, the receiver infers the sender's type through messages from the sender and the detector and then makes a decision. The timing of the game is as follows:

1. The designer designs the lie detector (the details are in the following paragraphs).
2. Nature draws the sender's type $\theta \in \{H, L\}$.
3. The sender sends a message $m \in \{m_H, m_L\}$ to the receiver.
4. The detector sends a signal $l \in \{a, na\}$ to the receiver.
5. The receiver takes an action $r \in \{r_H, r_L\}$.

Detector Design

A designer designs the detector. The designer's goal depends on the specific contexts, including maximizing the receiver's expected payoff, maximizing the high-type sender's expected payoff, and maximizing social welfare. We assume that the designer has access to an exogenously given classifier that generates a prediction for the message's trustworthiness when the sender sends message m_H.[12] The designer then

[11] A crucial feature of all signaling games is that the cost of sending a message depends on the sender's private type. Without this feature, the sender's message does not have a signaling role, and the problem becomes a cheap-talk game. Therefore, the positive lying cost C is an important ingredient of our model, though we allow it to be arbitrarily close to zero. If there is no lying cost, a low-type sender will always lie. This pooling equilibrium coincides with the main model only if both the false-positive and false-negative rates are low. In other cases, the equilibrium with a positive lying cost is semi-separating.

By choosing the detector, the designer can influence equilibrium outcomes through two channels: deterring the generation of disinformation and providing informative signals about the sender's message. If the lying cost is zero, the first channel is not available, though the designer can still affect the equilibrium outcome through the second channel.

[12] The exogenous classifier setup captures the idea that platforms or regulators face technology constraints and cannot perfectly classify messages. One can think of the endowed classifier as the best available one after the platforms or regulators try their best to improve the classifier's accuracy. Given this practical limitation, the designer still has some flexibility in deciding the rule of sending alarms, which affects the sender's equilibrium strategic information provision. Also, we will show that the sender must be the low type if he sends message m_L. So, there is no uncertainty about the sender's type when the receiver sees message m_L.

decides whether to send an alarm based on the prediction. More specifically, the classifier generates a binary outcome $s \in \{s_L, s_H\}$. We assume without loss of generality that the message is more likely to be truthful if the outcome is s_H. Formally, denote the probability of outcome s conditional on the sender's true type θ by $\phi(s|\theta)$. Then, $\phi(s_H|\theta = H) > \phi(s_H|\theta = L)$ and $\phi(s_L|\theta = L) > \phi(s_L|\theta = H)$.[13] We refer to ϕ as the *classifier's capacity*, as it reflects the quality of the classification. For example, the classifier perfectly reveals the truthfulness of the message if $\phi(s_H|\theta = H) = \phi(s_L|\theta = L) = 1$, whereas it is not very informative if $\phi(s_H|\theta = H)$ is close to $\phi(s_H|\theta = L)$ and $\phi(s_L|\theta = L)$ is close to $\phi(s_L|\theta = H)$. In reality, the classifier will not be perfect at classification because of practical limitations. So, we assume that $0 < \phi(s|\theta) < 1$ for $s \in \{s_H, s_L\}$ and $\theta \in \{H, L\}$.

Given the classifier's prediction, the designer decides whether to send an alarm. The designer's decision can be characterized by the probability of sending an alarm given classification outcome s_L, $\lambda_L = \Pr(l = a|s_L)$, and the probability of sending an alarm given classification outcome s_H, $\lambda_H = \Pr(l = a|s_H)$. We will refer to $\{\lambda_L, \lambda_H\}$ as the alarm rule. The detector may not always send an alarm given a negative signal s_L by the classifier because false-positive alarms can have significant economic and reputational consequences. So, even if there is some noisy evidence of disinformation, the designer does not necessarily want to label all such messages if they want to limit the false-positive rate. On the other hand, the detector may still want to sometimes send an alarm given a positive signal s_H by the classifier because the classifier generates noisy predictions and the message may still be deceptive. More frequent alarms can also serve as a threat, which reduces low-type senders' incentive of lying. This may happen if the designer really cares about lowering the amount of disinformation in equilibrium. In extension 5.1, we consider a restriction on the alarm rule where the detector never sends alarms when the classifier predicts signal s_H ($\lambda_H = 0$), and show that one can recover similar insights to the main model.

In a perfect world, a detector sends an alarm if and only if a low-type sender sends a deceptive message m_H. Therefore, the detector's **true-positive rate, denoted by** β, is the probability of sending an alarm when a type L sender sends message m_H, $\Pr(l = a \mid m = m_H, \theta = L)$. The detector will send a false alarm if a high-type sender sends a message m_H. So, the detector's **false-positive rate, denoted by** α, is $\Pr(l = a \mid m = m_H, \theta = H)$. For a given alarm rule $\{\lambda_L, \lambda_H\}$, the true-positive rate is $\beta = \phi(s_L|\theta = L)\lambda_L + \phi(s_H|\theta = L)\lambda_H$ and the false-positive rate is $\alpha = \phi(s_L|\theta = H)\lambda_L + \phi(s_H|\theta = H)\lambda_H$, according

[13] Conditions $\phi(s_H|\theta = H) > \phi(s_H|\theta = L)$ and $\phi(s_L|\theta = L) > \phi(s_L|\theta = H)$ are equivalent to $Pr(\theta = H|s_H) > Pr(\theta = H|s_L)$ by Bayes' rule.

to Bayes' rule. We refer to α and β as the *detector's capacity* because they reflect the quality of the detection. We can represent a detector by its capacity, $\{\beta, \alpha\}$, or by its alarm rule and the capacity of the classifier, $\{\lambda_L, \lambda_H, \phi\}$. Intuitively, a stronger detector correctly alarms a lie more frequently and mistakenly alarms a truth-telling message less frequently. This leads to the following definition.

Definition 1. A detector $\{\beta', \alpha'\}$ is stronger than a detector $\{\beta, \alpha\}$ if and only if the following conditions hold: $\beta' \geq \beta$, $\alpha' \leq \alpha$, and at least one of the inequalities is strict.

Receiver's Belief Updating

Without any information, the receiver's *prior belief* that the sender is high-type is ρ. The receiver updates her belief after the sender chooses a message based on the sender's strategy and Bayes' rule. The updated belief is the *intermediate belief* of the receiver. The receiver updates the belief again after observing the detector's signal. Because this belief takes into account all the available information the receiver can obtain, it is the *posterior belief* of the receiver. Figure 1 illustrates the receiver's belief updating processes.

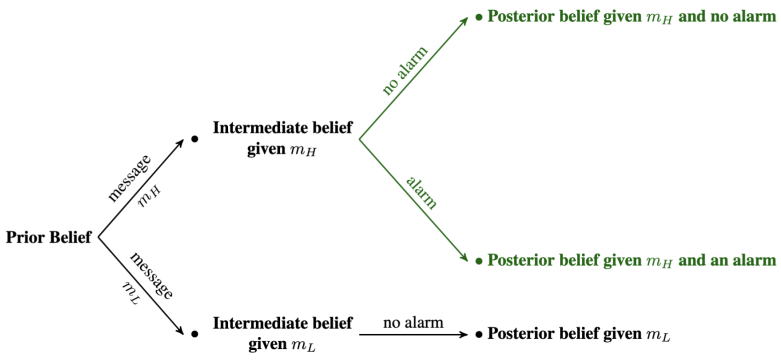


Figure 1: Receiver's Belief Updating Processes

In reality, the receiver may not observe the sender's message and the alarm signal sequentially. For example, they may see both the reviews and alarm information about a restaurant on Yelp. In such cases, the receiver will directly reach the posterior belief given the information. This does not change our analysis because we introduce the intermediate belief in order to present the intuition and underlying mechanisms, rather than to be interpreted literally.

2.2 Applications

Fake Reviews and Platform Detection

The rise of platforms such as Yelp and TripAdvisor has led to a vast accumulation of user reviews. Among these, prior research has documented the widespread phenomenon of fake reviews. Some low-quality firms (low-type senders) invest in generating fake reviews in an attempt to mislead consumers into purchasing their products. Aware of such manipulation, platforms (designers) have implemented various algorithms to detect fake reviews and assist consumers by labeling potentially deceptive content. Given the reviews and their labels, consumers (receivers) interpret the information and make decisions based on their resulting beliefs.

Email Marketing and Alerts

Email marketing has been steadily growing as a vital channel for reaching potential buyers because of its low cost. Sellers (senders) send marketing messages to addresses on their mailing lists, hoping the recipient (receiver) will click on the emails and make purchases. To protect users from low-quality or scam content, email service providers (designers) such as Outlook and Gmail have developed sophisticated algorithms to identify suspicious emails and issue alerts accordingly. Upon receiving an email, a user decides whether to click on it based on the available information, and email service providers must design appropriate alerting rules to maximize user welfare.

Social Media Content Moderation

Social media platforms rely heavily on high-quality user-generated content. However, content creators (senders) may sometimes post misleading information to boost engagement metrics, such as likes, shares, or click-through rates on embedded links. To preserve user experience and trust, platforms (designers) like Twitter and Facebook invest significantly in content moderation. When their algorithms detect potentially deceptive content, they may attach a warning label (alarm) to the post. Based on the content and the presence or absence of such a label, users (receivers) decide whether to engage with the post - by liking, sharing, or clicking on the link. Because false alarms on genuine content risk alienating creators, while failing to flag harmful content can mislead users, platforms must carefully design their moderation policies to balance these competing concerns.

2.3 Strategies and Equilibrium Concept

Because we are studying a multi-stage game with incomplete information, we consider the Perfect Bayesian Equilibrium (PBE hereafter). We denote the sender's strategy by $\sigma^S(m \mid \theta)$, the probability of sending message $m \in \{m_L, m_H\}$ when the sender's type is $\theta \in \{L, H\}$. We denote the receiver's posterior belief about the sender's type by $b(t \mid m, l)$, the probability that the sender's type is $t \in \{L, H\}$ given the sender's message $m \in \{m_L, m_H\}$ and the detector's signal $l \in \{a, na\}$. We denote the receiver's strategy by $\sigma^R(r \mid m, l)$, the probability that the receiver takes action $r \in \{r_L, r_H\}$ given the sender's message $m \in \{m_L, m_H\}$ and the detector's signal $l \in \{a, na\}$.

We can now represent the high-type sender's problem by $\max_{m\in\{m_H,m_L\}} \mathbf{1}_{[m=m_H]}[\sigma^R(r_H|m_H,a)\alpha + \sigma^R(r_H|m_H,na)(1-\alpha)]\Delta_H^S + \mathbf{1}_{[m=m_L]}[\sigma^R(r_H|m_L,a)\cdot 0 + \sigma^R(r_H|m_L,na)\cdot 1]\Delta_H^S$, the low-type sender's problem by $\max_{m\in\{m_H,m_L\}} \mathbf{1}_{[m=m_H]}[\sigma^R(r_H|m_H,a)\beta + \sigma^R(r_H|m_H,na)(1-\beta)]\Delta_L^S + \mathbf{1}_{[m=m_L]}[\sigma^R(r_H|m_L,a)\cdot 0 + \sigma^R(r_H|m_L,na)\cdot 1]\Delta_L^S$, and the receiver's problem by $\max_{r\in\{r_H,r_L\}} \mathbf{1}_{[r=r_H]}\cdot[b(H|m,l)\Delta_H^R - b(L|m,l)\Delta_L^R]$.

A PBE satisfies three properties: (1) Belief consistency: beliefs must be updated according to Bayes' rule along the equilibrium path; (2) Optimality: no player can improve their expected payoff by unilaterally deviating from their equilibrium strategy, given the strategies and beliefs of the other players; and (3) Sequential rationality: at every possible message and signal, the receiver's strategy must be optimal given her beliefs. Sequential rationality implies that the receiver will always choose action r_H if her posterior belief that the sender is of type H exceeds the threshold $\hat{\rho}$; she will always choose r_L if her posterior belief is below $\hat{\rho}$; and she may randomize between r_H and r_L if her posterior belief is exactly $\hat{\rho}$.

Intuitively, a high-type sender has no incentive to mimic the low type by sending costly disinformation. The following lemma formalizes the intuition that a type H sender always reports truthfully in equilibrium.

Lemma 1. *In any PBE, type H sender always sends the message $m = m_H$. The receiver always takes action $r = r_L$ after receiving message $m = m_L$, if the sender sends message m_L with a positive probability in equilibrium.*

To simplify notation, we denote the low-type sender's strategy by $\sigma^S \equiv \sigma^S(m_H \mid L)$ and denote the receiver's strategy by $\sigma_{na}^R \equiv \sigma^R(r_H \mid m_H, na), \sigma_a^R \equiv \sigma^R(r_H \mid m_H, a)$, and $\sigma_{L,na}^R \equiv \sigma^R(r_H \mid m_L, na)$. Because Lemma 1 has pinned down the strategy of type H sender and the strategy of the receiver upon receiving message m_L, in the subsequent analyses, we will use $\{\sigma^{S*}, \sigma_{na}^{R\,*}, \sigma_a^{R*}, \alpha^*, \beta^*\}$ to denote the entire equilibrium, and will use $\{\sigma^{S*}, \sigma_{na}^{R\,*}, \sigma_a^{R*}\}$ to denote the equilibrium with an exogenous lie detector.

3 Some Benchmarks

3.1 No alarm

Lie detection plays an important role in our model. To better understand its strategic role, we consider a benchmark where the detector never sends an alarm (both the true-positive and false-positive rates equal zero).

Lemma 2. *Suppose the detector always sends a no-alarm signal, na. In the unique PBE, the low-type sender sends message m_H with probability $\rho\Delta_H^R/[(1-\rho)\Delta_L^R] \in (0,1)$; the receiver has a posterior belief of $\hat{\rho}$ and takes action r_H with probability $C/\Delta_L^S \in (0,1)$ upon observing message m_H. Both the low-type sender and the receiver obtain zero expected payoff. The high-type sender obtains an expected payoff of $\Delta_H^S C/\Delta_L^S$.*

In this benchmark, the unique PBE is a semi-separating equilibrium. A high-type sender always sends a truthful message, whereas a low-type sender uses a mixed strategy, with some probability of sending a truthful message m_L and some probability of pretending to be the high-type by sending message m_H. The probability of lying is such that the receiver has a posterior belief of $\hat{\rho}$ upon seeing m_H, and is indifferent between taking either action. When the prior belief ρ is higher, the receiver is more inclined to believe that the sender is type H upon receiving message m_H for a given sender's strategy. Therefore, the low-type sender is more likely to mimic the high type. Upon receiving message m_H, the receiver does not know whether the sender is a truth-telling high type or a deceptive low type and uses a mixed strategy between actions r_H and r_L. The probability of taking action r_H is such that a low-type sender is indifferent between lying and truth-telling. The sender's cost of lying increases in C. For him to be indifferent between lying and not lying, the benefit of lying must also increase in C. So, in equilibrium, upon observing message m_H, the receiver takes the sender's desired action r_H more frequently when the lying cost C increases.

3.2 No false-positive alarm

Previous work on lie detection under strategic communication implicitly assumes that there is no false-positive alarm. Those studies only consider one type of mistake, where a detector may fail to send an alarm when there is disinformation (false negative). This section considers a benchmark consistent with previous literature by assuming that there is a detector that never sends a false-positive alarm ($\alpha = 0$).

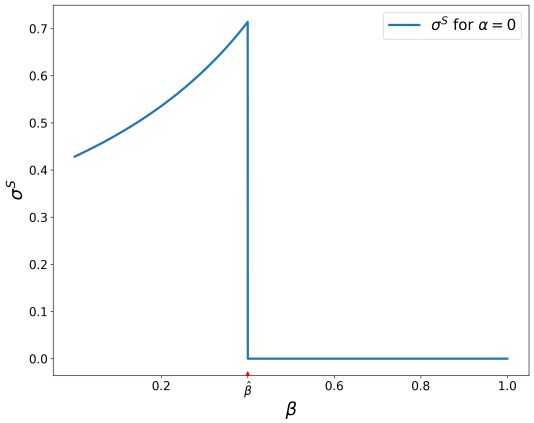


Figure 2: Probability of lying under different detectors without false-positive alarms for $\Delta_H^R = 0.5, \Delta_L^R = 0.5, \Delta_L^S = 0.5, C = 0.3, \rho = 0.3$, and any $\Delta_H^S > 0$.

Lemma 3 (Exogenous Detector)**.** *Consider a setting where the detector has zero false-positive rate. As the detector's true-positive rate β increases, the low-type sender exhibits a higher propensity to lie, but ceases lying when β exceeds the threshold $\hat{\beta} := 1 - C/\Delta_L^S$. The Perfect Bayesian Equilibria (PBEs) are characterized as follows:*

1. *For high lying cost $C \geq \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}\Delta_L^S$:*

$$\begin{cases} \sigma^{S*} = \frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, {\sigma_{na}^R}^* = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^{R*} = 0, & 0 < \beta < \hat{\beta} \\ \sigma^{S*} \in \left[0, \frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}\right], {\sigma_{na}^R}^* = 1, \sigma_a^{R*} = 0, & \beta = \hat{\beta} \\ \sigma^{S*} = 0, {\sigma_{na}^R}^* = 1, \sigma_a^{R*} \leq \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta > \hat{\beta} \end{cases}$$

2. *For low lying cost $C < \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}\Delta_L^S$:*

$$\begin{cases} \sigma^{S*} = \frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, {\sigma_{na}^R}^* = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^{R*} = 0, & 0 < \beta < 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \sigma^{S*} = 1, {\sigma_{na}^R}^* \in \left[\frac{C}{(1-\beta)\Delta_L^S}, 1\right], \sigma_a^{R*} = 0, & \beta = 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \sigma^{S*} = 1, {\sigma_{na}^R}^* = 1, \sigma_a^{R*} = 0, & 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} < \beta < \hat{\beta} \\ \sigma^{S*} \in [0,1], {\sigma_{na}^R}^* = 1, \sigma_a^{R*} = 0, & \beta = \hat{\beta} \\ \sigma^{S*} = 0, {\sigma_{na}^R}^* = 1, \sigma_a^{R*} \leq \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta > \hat{\beta} \end{cases}$$

Note that (m_L, na) is off the equilibrium path when $\sigma^{S*} = 1$, and (m_H, a) is off the equilibrium path when $\sigma^{S*} = 0$. The equilibrium beliefs and the strategy ${\sigma_{L,na}^R}^*$ satisfy the following conditions:

1. *When* $\sigma^{S*} = 1$*, the off-path belief* $b(H \mid m_L, na) \leq \hat{\rho}$*. When* $\sigma^{S*} = 0$*, the off-path belief* $b(H \mid m_H, a) \leq \hat{\rho}$*. In other cases, the beliefs are determined by Bayes' rule along the equilibrium path.*

2. *The strategy* ${\sigma^R_{L,na}}^*$ *satisfies:* ${\sigma^R_{L,na}}^* \leq (1-\beta){\sigma^R_{na}}^* - C/\Delta^S_L$ *if* $\sigma^{S*} = 1$*, and* ${\sigma^R_{L,na}}^* = 0$ *if* $\sigma^{S*} < 1$*.*

Figure 2 illustrates the low-type sender's equilibrium probability of lying as a function of the detector's true-positive rate β. In the absence of false-positive alarms, the receiver may see an alarm only if the sender is low-type. In that case, an alarm eliminates all the uncertainty about the sender's type. The receiver's posterior belief goes all the way to zero after observing an alarm. The receiver never takes the sender's desired action when there is an alarm. If a low-type sender is caught lying by the detector, he obtains no benefit from lying but instead incurs the cost of lying. When the true-positive rate is high, the expected payoff from lying is negative because the low-type sender has a high chance of being detected. So, the sender never lies and there is no disinformation when the true-positive rate β exceeds a threshold $\hat{\beta}$. We will show in the main model that this is not the case when we consider false-positive alarms.

We now study the endogenous detector design. To have a well-defined equilibrium payoff, we select the Pareto-optimal equilibrium for those cases with multiple equilibria. For a given detector $(\beta, 0)$, denote the receiver's expected equilibrium payoff by $\mathbb{E}U^R_0(\beta)$, the high-type sender's expected equilibrium payoff by $\mathbb{E}U^S_{0,H}(\beta)$, and the low-type sender's expected equilibrium payoff by $\mathbb{E}U^S_{0,L}(\beta)$. The social welfare is $\mathbb{E}W_0(\beta) = \mathbb{E}U^R_0(\beta) + \rho\mathbb{E}U^S_{0,H}(\beta) + (1-\rho)\mathbb{E}U^S_{0,L}(\beta)$. We consider three types of objectives by the designer, including choosing β to maximize the receiver's expected payoff $\mathbb{E}U^R_0(\beta)$, the high-type sender's expected payoff $\mathbb{E}U^S_{0,H}(\beta)$, or the social welfare $\mathbb{E}W_0(\beta)$.

Lemma 4 (Endogenous detector)**.** *The receiver's expected payoff, the high-type sender's expected payoff, and the social welfare all (weakly) increase in the true-positive rate* β*. The optimal true positive rate for the receiver is any* $\beta \geq \hat{\beta}$*. The optimal true positive rate for the high-type sender is any* $\beta \geq \min\{1 - \rho\Delta^R_H/[(1-\rho)\Delta^R_L], \hat{\beta}\}$*. The optimal true positive rate for social welfare is any* $\beta \geq \hat{\beta}$*.*

The receiver benefits from better distinguishing between the two types of senders, enabling more informed decision-making. As long as the true-positive rate exceeds $\hat{\beta}$, low-type senders stop lying because of the high likelihood of being caught. So, the receiver can perfectly infer a sender's type from the sender's message m and attains the highest possible payoff for any sufficiently high β. There is never an alarm when the sender is high type. So, a high-type sender only cares about the receiver's action upon seeing no alarm.

When $\beta \geq \min\{1 - \rho\Delta_H^R/[(1-\rho)\Delta_L^R], \hat{\beta}\}$, the receiver always takes the sender's desired action, allowing the high-type sender to achieve the highest payoff. Because society benefits from a lower level of disinformation, social welfare reaches its maximum when β is sufficiently high to deter the low-type sender from lying.

The general message from the lemma is simple and intuitive: when there is no false positive, the more accurate the detector is, the better. So, there is no trade-off, and the detector designer always prefers a higher true-positive rate. In reality, the detector may make another type of mistake by sending a false alarm in the absence of disinformation (false positive). It is generally impossible to eliminate either type of mistake unless the detector always or never sends alarms. We will show in the main model that the designer strictly prefers an intermediate true-positive rate to the highest true-positive rate in the presence of false-negative alarms. A higher true-positive rate may reduce the receiver's expected payoff, the high-type sender's expected payoff, and the social welfare. In this case, the optimal detector design becomes both non-trivial and managerially important.

4 Equilibrium

4.1 Equilibrium with an exogenous detector

We first consider the equilibrium with an exogenous detector $\{\beta, \alpha\}$ for four reasons. First, by abstracting away the strategic role of the designer/detector, we can more clearly understand the driving force behind different results. Second, previous literature focuses on the case with an exogenous detector. This section enables us to compare the equilibrium outcomes cleanly with the benchmark of no false positives. Third, the entire equilibrium is complicated by the presence of three strategic players: the sender, the receiver, and the designer/detector. The equilibrium with only the sender and the receiver is simpler to solve and serves as a building block to solve the entire equilibrium. Last, it applies to scenarios where a platform or regulator adopts a given detector rather than adjusting or developing one.

The detector is uninformative if $\alpha = \beta$. In such cases, the equilibrium outcome is essentially the same as the equilibrium of the no-alarm benchmark in section 3.1.[14] We will present the formal characterization of the equilibrium in the appendix, and will show that an uninformative detector is never optimal in equilibrium.

[14]Technically, there is a subtle difference between the two cases because we need to specify the receiver's strategies both upon receiving an alarm and upon receiving no alarm in the $\alpha = \beta > 0$ case, though either strategy is the same as the receiver's strategy in the no alarm benchmark because the detector is not informative.

For any detector such that $\alpha > \beta$, we can obtain essentially the same equilibrium outcome with an alternative detector whose $\alpha < \beta$. Therefore, we focus on the interesting case where the detector may send both types of alarms and the false-positive rate is lower than the true-positive rate, $0 < \alpha < \beta$.

Proposition 1 (Equilibrium with an Exogenous Detector)**.** *Suppose the detector $\{\beta, \alpha\}$, $0 < \alpha < \beta$, is exogenously given. The receiver's posterior belief about the sender being type H upon observing message m_H and a noisy signal $l \in \{n, na\}$ is the following.*

$$b(H \mid m_H, l) = \begin{cases} \frac{\alpha\rho}{\alpha\rho+\beta\sigma^S(1-\rho)}, & l = a \\ \frac{(1-\alpha)\rho}{(1-\alpha)\rho+(1-\beta)\sigma^S(1-\rho)}, & l = na \end{cases}$$

The low-type sender's probability of lying first increases and then decreases in the detector's true-positive rate β. The PBEs are the following.

$$\sigma^{S*} \begin{cases} = \min\left\{\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, 1\right\}, & \beta < \hat{\beta} \\ \in \left[\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \min\left\{\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, 1\right\}\right], & \beta = \hat{\beta} \\ = \frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, & \beta > \hat{\beta} \end{cases}$$

$${\sigma_{na}^R}^* \begin{cases} = 1, & \beta > 1 - \frac{(1-\alpha)\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \in \left[\min\left\{\frac{C}{(1-\beta)\Delta_L^S}, 1\right\}, 1\right], & \beta = 1 - \frac{(1-\alpha)\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ = \min\left\{\frac{C}{(1-\beta)\Delta_L^S}, 1\right\}, & \beta \in (\alpha, 1 - \frac{(1-\alpha)\rho\Delta_H^R}{(1-\rho)\Delta_L^R}) \end{cases}, \quad \sigma_a^{R*} = \max\left\{\frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, 0\right\}$$

Note that (m_L, na) is off the equilibrium path when $\sigma^{S*} = 1$. The equilibrium beliefs and strategy ${\sigma_{L,na}^R}^*$ satisfy the following conditions:

1. *When $\sigma^{S*} = 1$, the off-path belief $b(H \mid m_L, na) \leq \hat{\rho}$. In other cases, the beliefs are determined by Bayes' rule along the equilibrium path.*
2. *The strategy ${\sigma_{L,na}^R}^*$ satisfies: ${\sigma_{L,na}^R}^* \leq (1-\beta){\sigma_{na}^R}^* - \frac{C}{\Delta_L^S}$ if $\sigma^{S*} = 1$, and ${\sigma_{L,na}^R}^* = 0$ if $\sigma^{S*} < 1$.*

Table 2 summarizes the equilibrium strategy $(\sigma^S, \sigma_{na}^R, \sigma_a^R)$.

There are three categories of equilibria in a signaling game: pooling equilibrium, separating equilibrium, and semi-separating equilibrium. We will focus on the semi-separating equilibrium in this section because it is the non-trivial case and the only type of equilibrium outcome under most parameter ranges. It is also

β Range \ α Range	$\alpha < 1 - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta)$	$\alpha = 1 - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta)$	$\alpha \in \left(1 - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta), \beta\right)$	$\alpha = \beta$
$\beta \in \left(\hat{\beta}, 1\right]$	$\sigma^S = \frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}$			$\sigma^S = \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}$,
$\beta = \hat{\beta}$	$\sigma^S \in \left[\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \min\left\{\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, 1\right\}\right], \sigma_{na}^R = 1, \sigma_a^R = 0$			$\sigma_{na}^R, \sigma_a^R$
$\beta \in \left(0, \hat{\beta}\right)$	$\sigma^S = 1,$ $\sigma_{na}^R = 1,$ $\sigma_a^R = 0$	$\sigma^S = 1,$ $\sigma_{na}^R \in \left[\frac{C}{(1-\beta)\Delta_L^S}, 1\right],$ $\sigma_a^R = 0$	$\sigma^S = \frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}$ $\sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S},$ $\sigma_a^R = 0$	such that $\beta\sigma_a^R + (1-\beta)\sigma_{na}^R$ $= \frac{C}{\Delta_L^S}$

Table 2: Equilibria with an exogenous detector

the case that is most relevant to the endogenous detector design problem, where many detectors that induce other types of equilibrium are not feasible.

4.1.1 Effect of Lie Detection on Receiver's Posterior Belief

After observing message m_H but before observing the detector's signal, the receiver's intermediate belief about the sender's type is $\Pr(\theta = H | m = m_H) = \rho/[\sigma^S(1-\rho)+\rho]$. We now disentangle two effects of lie detection on the receiver's posterior belief, illustrated by Figure 3. We also examine how a stronger lie detector changes the belief-updating process by comparing the receiver's posterior beliefs under detector $\{\beta, \alpha\}$ and a stronger detector $\{\beta', \alpha'\}$.

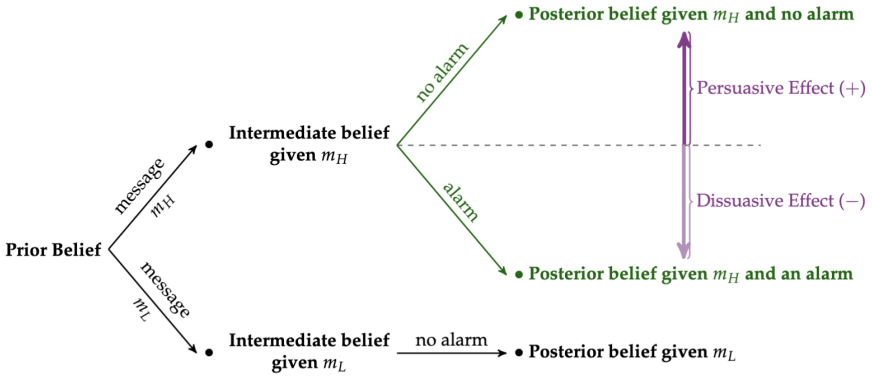


Figure 3: The Effect of Lie Detection on the Receiver's Belief.

1. Persuasive effect: Because the detector is more likely to send no alarm when the sender is high-type than when he is low-type, the receiver becomes more certain that the sender is high-type if she receives no alarm. The presence of a detector persuades the receiver to trust the sender's m_H message more in this case. We call this posterior belief-enhancing effect a persuasive effect. Formally, the persuasive

effect raises the receiver's belief from the intermediate belief, $\Pr(\theta = H|m = m_H) = \rho/[\sigma^S(1-\rho)+\rho]$, to the posterior belief, $\Pr(\theta = H|m = m_H, l = na) = (1-\alpha)\rho/[(1-\alpha)\rho+(1-\beta)\sigma^S(1-\rho)]$. The posterior belief increases in the true-positive rate and decreases in the false-positive rate. Therefore, *the persuasive effect is larger under a stronger detector*, as illustrated by Figure 4.

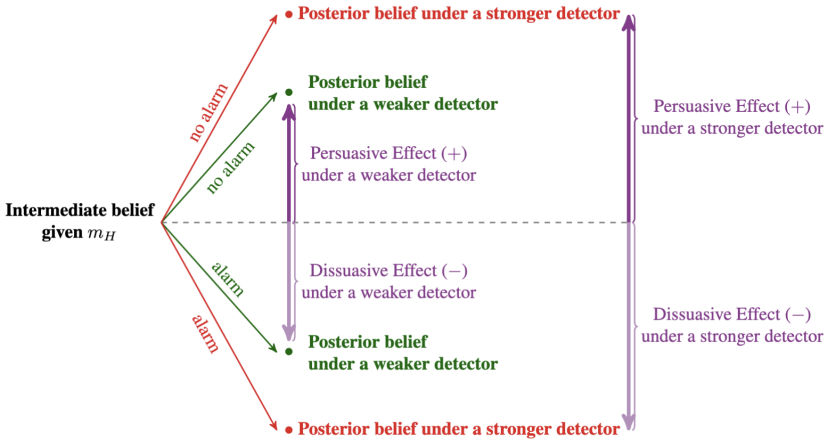


Figure 4: The Effect of a Stronger Detector on the Receiver's Belief.

2. Dissuasive effect: Because the detector is more likely to send an alarm when the sender is low-type than when he is high-type, the receiver becomes more certain that the sender is low-type if she receives an alarm. The presence of an alarm makes the receiver less trustful about the sender's m_H message. We call this posterior belief-reducing effect a dissuasive effect. Formally, the dissuasive effect reduces the receiver's belief from the intermediate belief, $\Pr(\theta = H|m = m_H) = \rho/[\sigma^S(1-\rho)+\rho]$, to the posterior belief, $\Pr(\theta = H|m = m_H, l = a) = \alpha\rho/[\alpha\rho+\beta\sigma^S(1-\rho)]$. The posterior belief decreases in the true-positive rate and increases in the false-positive rate. Therefore, *the (absolute value of the) dissuasive effect is larger under a stronger detector*, as illustrated by Figure 4.

   A key difference between our setting and the no false-positive alarm benchmark is related to the dissuasive effect. In the absence of false-positive alarms ($\alpha = 0$), the receiver may see an alarm only if the sender is low-type. Therefore, an alarm eliminates all the uncertainty about the sender's type. The receiver's posterior belief goes all the way to zero after observing an alarm. Thus, the dissuasive effect does not depend on the true-positive rate of the detector; two detectors with very different β generate the same effect on the posterior belief if they send an alarm. In contrast, in the presence of false-positive alarms, two detectors with the same false-positive rate but different true-positive rates generate different dissuasive effects. As we will show in the next subsection, variations in the

dissuasive effects lead to qualitatively different equilibrium outcomes.

4.1.2 Non-monotonic Relationship Between Detector's Capacity and Sender's Probability of Lying

According to Proposition 1, there is a non-monotonic relationship between the detector's capacity α and β and a low-type sender's probability of lying σ^S. Figure 5 illustrates such non-monotonicity by plotting a low-type sender's probability of lying as a function of the detector's true-positive rate for three fixed false-positive rates. As we can see from the figure, a stronger detector increases the probability of lying when the true-positive rate is low and decreases the probability of lying when the true-positive rate is high. Below, we discuss the underlying mechanism and intuition in detail.

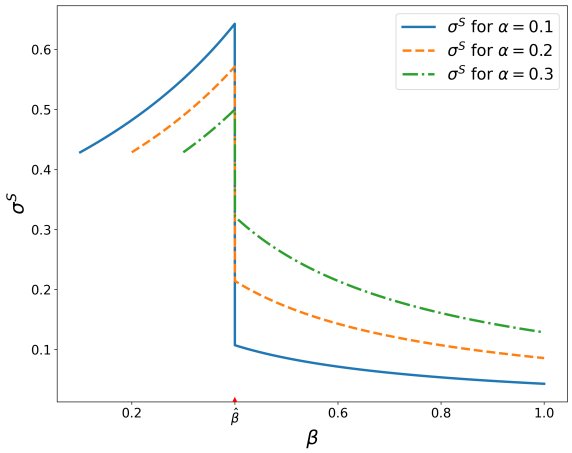


Figure 5: Probability of lying under different detectors for $\Delta_H^R = 0.5, \Delta_L^R = -0.5, \Delta_L^S = 0.5, C = 0.3, \rho = 0.3$, and any $\Delta_H^S > 0$.

Low True-positive Rate

When the detector's true-positive rate β is low, the detector will fail to catch many low-type senders who are lying. This creates a strong incentive for a low-type sender to pretend to be a high-type. So, the probability of lying is at a relatively high level. This leads to a low posterior belief. Therefore, the receiver will never take the sender's desired action upon observing an alarm.

If the receiver always takes the sender's desired action r_H after observing message m_H and no alarm, the expected benefit of lying is $(1-\beta)\Delta_L^S$, which is larger than the lying cost C when β is low. However, this implies that a low-type sender will always lie. The high likelihood of disinformation leads to a low intermediate belief given message m_H, which indicates a low posterior belief even without an alarm. Therefore, the receiver will not take action r_H upon observing no alarm. This is a contradiction. If the receiver never takes the sender's desired action r_H after observing message m_H and no alarm, no low-type sender will lie.

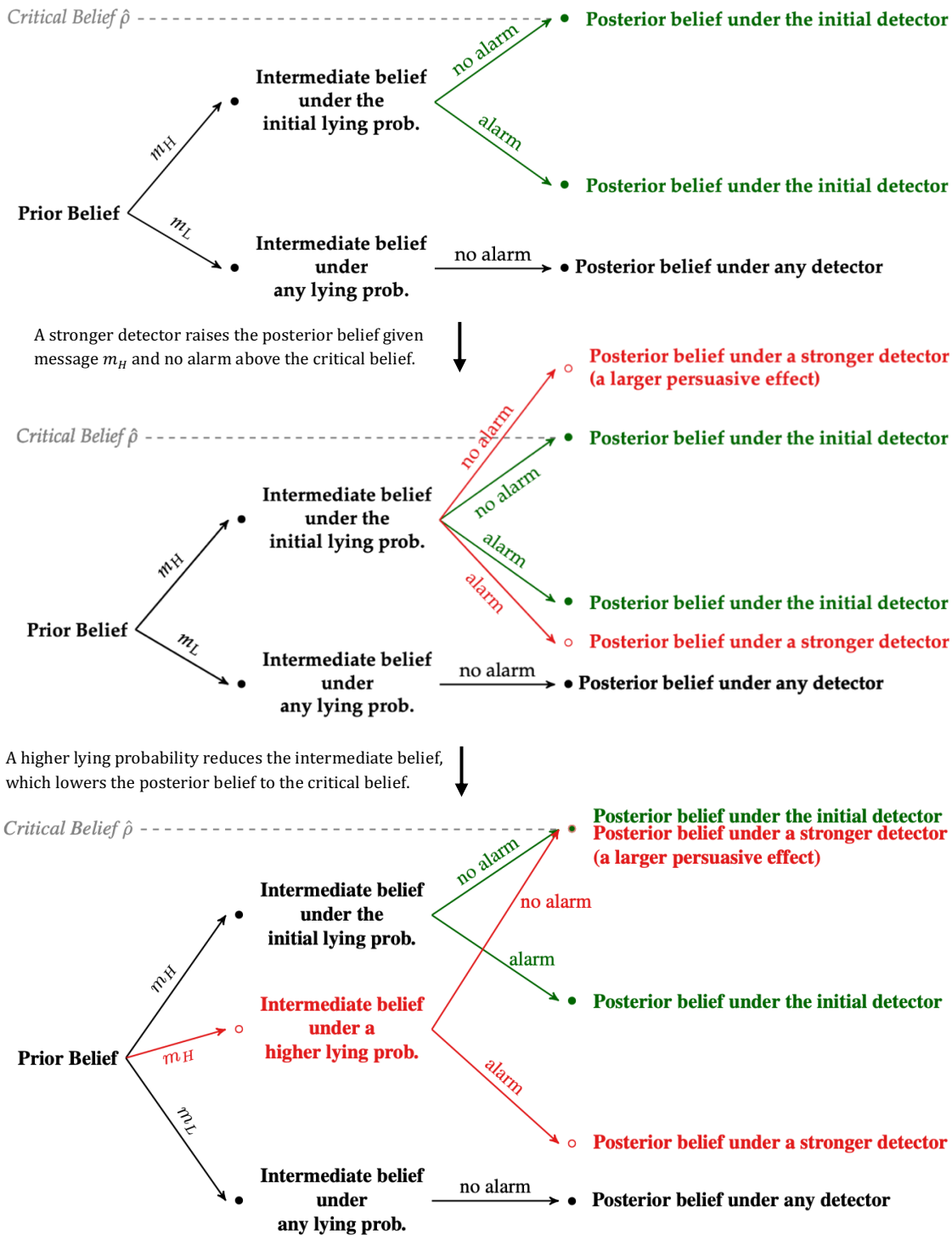


Figure 6: How a stronger detector increases the probability of lying when β is low.

But this implies that only the high type will send message m_H and that the receiver's posterior belief upon observing m_H will be one. Hence, the receiver should always take action r_H upon observing message m_H. This is also a contradiction. In sum, the only equilibrium for the receiver is to follow a mixed strategy after observing message m_H and no alarm. This implies that her posterior belief after observing message m_H and no alarm must be exactly $\hat{\rho}$.

Suppose the lying probability does not change, the receiver's posterior belief after seeing no alarm will exceed $\hat{\rho}$ under a stronger detector because of a larger persuasive effect. To keep the belief at $\hat{\rho}$, the equilibrium lying probability must be higher, which lowers the base rate of high-type senders conditional on sending message m_H. This reduces the intermediate belief, and thus the posterior belief. Therefore, low-type senders lie more frequently as the detector becomes stronger. Figure 6 illustrates the mechanism.

High True-positive Rate

When the detector's true-positive rate β is high, the detector will catch a high proportion of low-type senders who are lying. This creates a low incentive for a low-type sender to pretend to be a high type. So, the probability of lying is at a relatively low level. This implies that the base rate of a low-type sender conditional on sending message m_H is low. So, after observing message m_H and before observing the alarm, the receiver has a high intermediate belief about the sender being high-type. Consequently, the receiver always takes the sender's desired action if there is no alarm. Additionally, even after the receiver observes an alarm that reduces her belief, the posterior belief is still high enough such that the receiver may take the sender's desired action with a positive probability.

Suppose the receiver's posterior belief after observing an alarm is higher than $\hat{\rho}$. The receiver always takes the sender's desired action regardless of the alarm. Then, a low-type sender will always lie because the benefit of lying Δ_L^S is larger than the lying cost C. In this scenario, the prior belief equals the intermediate belief because all senders send the same message. But then the receiver's posterior belief after observing an alarm, which is always lower than the intermediate belief, will be lower than the prior belief $\rho < \hat{\rho}$. This is a contradiction. Suppose, instead, that the posterior belief after observing an alarm is lower than $\hat{\rho}$. The receiver will never take action r_H after observing an alarm. A low-type sender pretending to be a high type will not be detected with probability $1 - \beta$. Even if the receiver always takes the sender's desired action upon receiving no alarm, the expected benefit of lying is $(1 - \beta)\Delta_L^S$, which is smaller than the lying cost C when β is high. So, no low-type sender will lie. However, in that case, an alarm can only be a false-positive

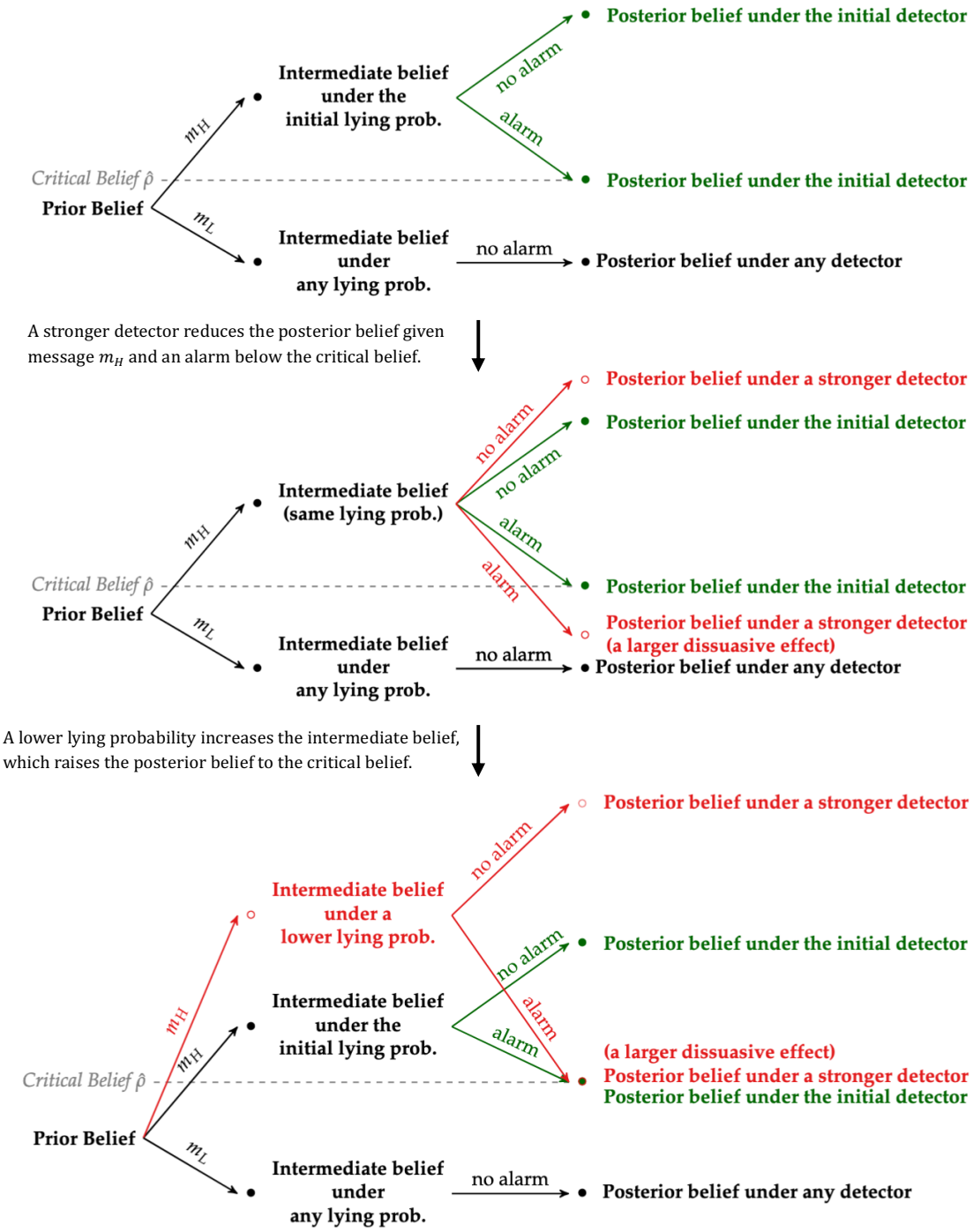


Figure 7: How a stronger detector decreases the probability of lying when β is high.

alarm, and the receiver's posterior belief after seeing message m_H will be one regardless of the alarm signal. This is also a contradiction. In sum, the receiver's posterior belief after observing an alarm must be exactly $\hat{\rho}$. This implies that the receiver adopts a mixed strategy after observing message m_H and an alarm.

Suppose the lying probability does not change, the receiver's posterior belief after observing an alarm will fall below $\hat{\rho}$ because of a larger dissuasive effect. To maintain the posterior belief at $\hat{\rho}$, the equilibrium lying probability must be lower, which raises the base rate of high-type senders conditional on sending message m_H. This raises the intermediate belief, and thus the posterior belief. Therefore, low-type senders lie less frequently as the detector becomes stronger. Figure 7 illustrates the mechanism. This decreasing pattern of the probability of lying is absent in the no-false-positive benchmark because there is no variation in the dissuasive effect when $\alpha = 0$: a stronger detector with a higher β does not affect the receiver's inference about the sender's type conditional on seeing an alarm.

Note that in the main model, even though the equilibrium lying probability reduces as the true-positive rate increases, it remains positive even when the true-positive rate is very high. This is because an alarm must be a false positive if no sender is lying. So, the receiver's belief after seeing message m_H will be one in such a separating equilibrium, regardless of the alarm. By deviating from truth-telling, a low-type sender will never be caught and will always mislead the receiver into taking the sender's desired action. Therefore, the no-lying equilibrium cannot be sustained, and there will always be some disinformation in equilibrium.

The Cut-off Threshold $\hat{\beta}$

Figure 5 and its discussion illustrated that low-type senders adopt markedly different strategies depending on whether the true-positive rate β is low or high. The cut-off threshold for the true-positive rate is given by $\hat{\beta} = 1 - C/\Delta_L^S$. When the benefit of successfully mimicking high-type senders, Δ_L^S, is greater, low-type senders have a stronger incentive to imitate the high type. As a result, they lie frequently across a wider range of β, and the threshold $\hat{\beta}$ increases with Δ_L^S. Conversely, when the cost of lying C is higher, the incentive to mimic diminishes. Low-type senders are then more easily deterred from lying and lie frequently only within a narrower range of β, leading to a decrease in the threshold $\hat{\beta}$.

In sum, persuasive and dissuasive effects jointly drive the non-monotonic relationship between the detector's capacity and the sender's probability of lying. The persuasive effect leads to an increasing pattern when the detector is weak, whereas the dissuasive effect generates a decreasing pattern when the detector is strong.

4.1.3 Discussion on the Equilibrium Type

No separating equilibrium exists in the main model. In contrast, in the no-false-positive benchmark, there is a separating equilibrium when the true-positive rate is high. The key economic force driving the different outcomes is whether the receiver may take the sender's desired action even after observing an alarm. Without false-positive alarms, the receiver knows with certainty that the sender is lying when they see an alarm. In such cases, the receiver never takes the sender's desired action, and the low-type sender always obtains a negative utility because of the lying cost. When the true-positive rate is high, disinformation is detected at a high rate, deterring a low-type sender from lying. Therefore, the equilibrium is separating - a high type always sends message m_H and a low type always sends message m_L. With false-positive alarms, Section 4.1.2 shows that the receiver may take the sender's desired action even after observing an alarm. Such a possibility implies that a low-type sender will deviate from the separating equilibrium because an alarm can only be a false-positive one in such an equilibrium, and thus the low type can always mimic the high type successfully. Type I and type II errors may be viewed conceptually as similar to one another in that both types of errors make information less precise. The stark difference in equilibrium outcomes with and without type I error indicates that false positives and false negatives generate different effects on strategic communication. In particular, the noise introduced by Type I error creates sufficient incentive for some low-type senders to mimic the high type, eliminating the possibility of separating equilibria. While Type II error also generates noisy information, it is insufficient on its own to preclude the existence of a separating equilibrium.

There is a pooling equilibrium in the main model only if both the false-positive and the true-positive rates are low. Section 4.1.1 shows that a low false-positive rate leads to a stronger persuasive effect, increasing the receiver's posterior belief after seeing no alarm. Consequently, the receiver always takes the sender's desired action when the sender sends message m_H and there is no alarm. A low true-positive rate implies that the detector will only catch a small proportion of disinformation. So, even if a lying low type obtains a positive payoff only when the detector fails to detect the lie, it happens with a high probability, justifying the lying cost. When both conditions hold, the low-type sender always lies. In all other cases, a pooling equilibrium does not exist.

4.2 Entire equilibrium

We now study the entire equilibrium where the detector is endogenously determined. Proposition 1 has characterized the sender's and the receiver's strategies for any given detector. So, we only need to determine the designer's strategy. It is challenging to determine the optimal detector because of the large number of detectors. Instead of choosing the true-positive and false-positive rates of the detector simultaneously, the next section suggests that we can simplify the optimization problem by pinning them down sequentially.

4.2.1 Effect of Lie Detection on the Payoffs

We need a well-defined equilibrium payoff to study the effect of lie detection on the payoffs. According to Proposition 1, the equilibrium is unique as long as $\alpha \neq 1 - [(1-\rho)\Delta_L^R/(\rho\Delta_H^R)](1-\beta), \beta \neq \hat{\beta}$, and $\alpha \neq \beta$. For those cases with multiple equilibria, we select the Pareto-optimal equilibrium.[15] Section A.2 in the appendix contains details of the refinement. The next proposition summarizes the effect of lie detection on the expected payoff of the receiver, low-type sender, and high-type sender.

Proposition 2. *1. For a given true-positive rate β, the expected payoff of the receiver, the expected payoff of the low-type sender, and the expected payoff of the high-type sender are all weakly decreasing in the false-positive rate α.*

2. For a given false-positive rate α, the expected payoff of the receiver and the expected payoff of the high-type sender are weakly increasing in the true-positive rate β, and the expected payoff of the low-type sender is weakly decreasing in the true-positive rate β.

The receiver benefits from a more informed decision. She wants to better match the action with the sender's true type. A stronger detector has a direct effect and an indirect effect on the receiver's posterior belief. For a given intermediate belief, the larger persuasive and dissuasive effects indicate that the receiver's signal is Blackwell more informative under a stronger detector. It also has an indirect effect, affecting low-type senders' lying probability. When the true-positive rate is high, a stronger detector reduces the lying probability, thereby further strengthening the receiver's ability to distinguish between high- and low-type senders. When the true-positive rate and the false-positive rate are low, the sender's lying probability does

[15] This refinement ensures that for any given detector (β, α), there is a unique equilibrium, thereby a unique equilibrium payoff. So, we can compare the payoff of different detectors when we determine the endogenous detector. The refinement does not drive the results on payoffs or welfare because the area in the detector's capacity space $\{(\beta,\alpha)|0 < \alpha \leq \beta \leq 1\}$ with multiple equilibria (for an exogenous detector), $\{(\beta,\alpha)|\alpha = 1 - [(1-\rho)\Delta_L^R/(\rho\Delta_H^R)](1-\beta)$ *or* $\beta = \hat{\beta}$ *or* $\alpha = \beta\}$, has measure zero.

not depend on the detector. When the true-positive rate is low and the false-positive rate is high, a stronger detector induces a higher probability of lying, making high- and low-type senders less distinguishable. In this case, the direct effect and the indirect effect of a stronger detector are opposite to each other. However, the direct effect dominates the indirect effect. In particular, consider a detector $\{\beta, \alpha\}$ in the parameter range of this case. As explained in Section 4.1.2 and Figure 6, the receiver's posterior belief after seeing no alarm is always $\hat{\rho}$. According to Proposition 1, the receiver's posterior belief after seeing an alarm is $\alpha\rho/\{\alpha\rho + \beta(1-\alpha)\rho\Delta_H^R/[(1-\beta)\Delta_L^R]\}$, which increases in α and decreases in β. Thus, even if we take into account the sender's changing equilibrium lying behavior, the receiver's signal becomes Blackwell more informative (the same posterior belief when there is no alarm and a lower posterior belief when there is an alarm) if we replace a detector with a stronger one. In other words, a stronger detector can better separate senders. Therefore, the receiver's payoff increases in the true-positive rate and decreases in the false-positive rate. Similarly, a high-type sender wants to distinguish himself from a low-type sender and, therefore, benefits from a higher true-positive rate and a lower false-positive rate.

The effect of lie detection on a low-type sender's payoff is more complicated. A low-type sender benefits from successfully pretending to be a high-type sender. A higher true-positive rate raises the likelihood that the low-type sender will be caught by the detector. So, a low-type sender's payoff decreases in the true-positive rate. Interestingly, a low-type sender's payoff also decreases in the false-positive rate, though a higher false-positive rate makes it harder to distinguish between the two types. The reason is as follows. A low-type sender obtains a positive payoff only if there is no alarm and the receiver takes action r_H in the absence of an alarm. A higher false-positive rate does not reduce the low-type sender's likelihood of being detected. However, it leads to a smaller persuasive effect when there is no alarm (please refer to Figure 4 and the discussion in Section 4.1.1). The receiver's posterior belief after observing no alarm decreases in α, and the receiver uses a mixed strategy rather than always taking action r_H when the belief hits $\hat{\rho}$; this hurts the low-type sender's payoff.

As one can see from the proposition, a lower false-positive rate makes all players better off. Thus, the designer always chooses the *lowest feasible false-positive rate* for any true-positive rate.

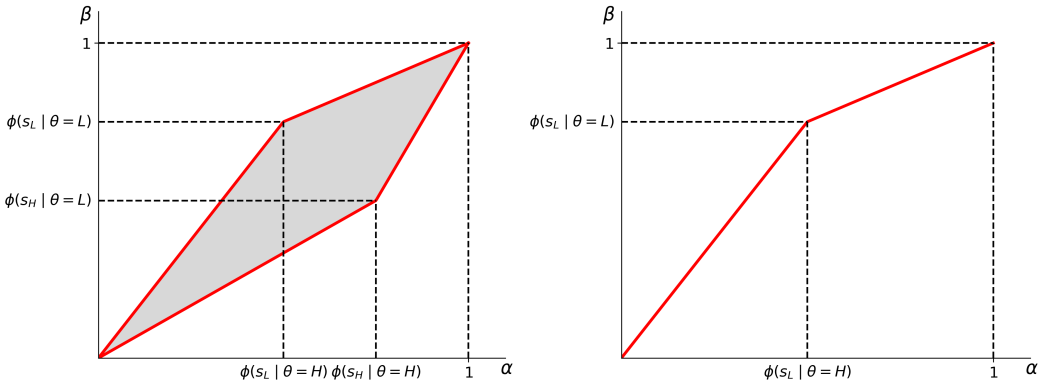
4.2.2 Optimal False-positive Rate and Alarm Rule Given True-positive Rate

Because of the constraint of the classifier's capacity, the designer cannot obtain all detectors $(\beta, \alpha) \in \{(\beta, \alpha) | 0 \leq \alpha < \beta \leq 1\}$ by choosing an alarm rule $\{\lambda_H, \lambda_L\}$. In particular, the space of the feasible

detectors given a classifier ϕ is $\mathcal{F}(\phi) := \{(\beta, \alpha) | \beta = \phi(s_L|\theta = L)\lambda_L + \phi(s_H|\theta = L)\lambda_H, \alpha = \phi(s_L|\theta = H)\lambda_L + \phi(s_H|\theta = H)\lambda_H, \lambda_L \in [0,1], \lambda_H \in [0,1]\}$, illustrated by Figure 8a.

Figure 8b presents the receiver operating characteristic curve (ROC curve), which represents the Pareto frontier of the classification outcome. The optimal detector must be a point on the ROC curve. According to the previous section, the ROC curve can be pinned down by choosing the *lowest feasible false-positive rate* for any true-positive rate. As we can see, the false-positive rate increases in the true-positive rate. The designer needs to sacrifice one metric in order to improve the other. Furthermore, the false-positive rate increases in the true-positive rate at a lower rate when β is low and at a higher rate when β is high.

The next result characterizes the optimal false-positive rate and alarm rule for a given true-positive rate.



(a) Feasible detectors $\mathcal{F}(\phi)$ (shaded region) (b) Receiver operating characteristic (ROC) curve

Figure 8: Detector design

Lemma 5 (Optimal False-positive Rate and Alarm Rule Given True-positive Rate)**.** *For a given true-positive rate β, the detector's optimal false-positive rate, denoted by $\alpha^*(\beta;\phi)$, is*

$$\alpha^*(\beta;\phi) = \begin{cases} \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\beta, & if\ \beta \leq \phi(s_L|\theta = L) \\ \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\beta + 1 - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}, & if\ \beta > \phi(s_L|\theta = L), \end{cases}$$

which increases in β. The detector $\{\beta, \alpha^*(\beta;\phi)\}$ can be achieved by the alarm rule:

$$\lambda_L^*(\beta) = \begin{cases} \frac{\beta}{\phi(s_L|\theta=L)}, & if\ \beta \leq \phi(s_L|\theta = L) \\ 1, & if\ \beta > \phi(s_L|\theta = L) \end{cases}, \ \lambda_H^*(\beta) = \begin{cases} 0, & if\ \beta \leq \phi(s_L|\theta = L) \\ \frac{\beta - \phi(s_L|\theta=L)}{\phi(s_H|\theta=L)}, & if\ \beta > \phi(s_L|\theta = L). \end{cases}$$

When choosing the alarm rule, the designer wants to achieve a given true-positive rate while minimizing

the false-positive rate. Intuitively, it is better to send an alarm in scenarios where the likelihood of the message being deceptive is higher. Because the sender is more likely to be a low type under signal s_L than under signal s_H, the detector sends fewer false alarms, conditional on sending the same amount of true alarms, by sending alarms after getting prediction s_L rather than s_H from the classifier. As a result, the designer prefers sending an alarm after getting prediction s_L. To achieve a low true-positive rate, the detector does not need to send any alarms after getting prediction s_H. So, $\lambda_H^*(\beta) = 0$ and $\lambda_L^*(\beta)$ increases in β for low β, as illustrated by Figure 9. Because each alarm falsely recognizes a high-type sender as a low-type with some probability, the false-positive rate also increases in β.

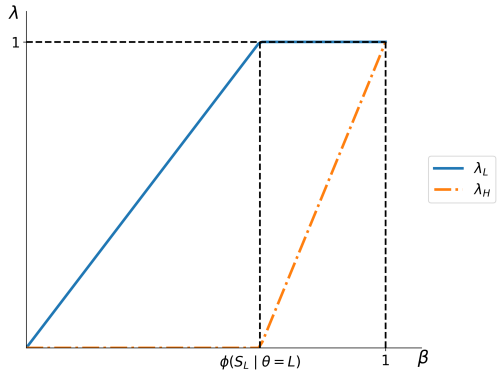


Figure 9: Optimal alarm rule for a given true-positive rate

The true-positive rate is capped by $\phi(s_L|\theta = L)$ even if the detector always sends an alarm after getting prediction s_L. So, the detector must also sometimes send an alarm after getting prediction s_H to achieve a true-positive rate above $\phi(s_L|\theta = L)$. In such cases, $\lambda_L^*(\beta) = 1$ and $\lambda_H^*(\beta)$ increases in β. In addition, the likelihood of the sender being a low type is smaller given prediction s_H rather than s_L. Compared to the low β case, the detector needs to raise the alarm probability by a larger value to add a unit to the true-positive rate. Consequently, $\lambda_H^*(\beta)$ and $\alpha^*(\beta;\phi)$ increase in β in this case at a higher rate than do $\lambda_L^*(\beta)$ and $\alpha^*(\beta;\phi)$ in the low β case.

4.2.3 Optimal Design of Lie Detector

We now study the detector designer's equilibrium strategy, which is the optimal choice of a feasible detector, $\{\beta^*, \alpha^*\}$. Lemma 5 has shown that the optimal false-positive rate is $\alpha^*(\beta;\phi)$ given any true-positive rate β. So, we only need to pin down the optimal true-positive rate β^*. For a given detector, denote the receiver's expected equilibrium payoff by $\mathbb{E}U^R(\beta,\alpha)$, the high-type sender's expected equilibrium payoff

by $\mathbb{E}U_H^S(\beta,\alpha)$, and the low-type sender's expected equilibrium payoff by $\mathbb{E}U_L^S(\beta,\alpha)$. The following lemma assures that sequential derivation of the optimal detector is equivalent to deriving β^* and α^* simultaneously.

Lemma 6. *Suppose the designer wants to choose a feasible detector to maximize* $J(\beta,\alpha) := w_R\mathbb{E}U^R(\beta,\alpha)+ w_H\mathbb{E}U_H^S(\beta,\alpha) + w_L\mathbb{E}U_L^S(\beta,\alpha)$, *where* $w_R, w_H, w_L \geq 0$ *with at least one inequality strict. The set of detectors that maximizes the designer's objective function among all detectors with the lowest feasible false-positive rate for a given true-positive rate is identical to the set of detectors that maximizes the designer's objective function among all feasible detectors,* $\{(\beta,\alpha^*(\beta;\phi))|J(\beta,\alpha^*(\beta;\phi)) = \max_\beta J(\beta,\alpha^*(\beta;\phi)), 0 < \beta \leq 1\} = \{(\beta,\alpha)|J(\beta,\alpha) = \max_{\beta,\alpha} J(\beta,\alpha), (\beta,\alpha) \in \mathcal{F}(\phi)\}$.

The classifier is more likely to generate outcome s_H if the sender's type is H rather than L, $\phi(s_H|\theta = H) > \phi(s_H|\theta = L)$, and more likely to generate outcome s_L if the sender's type is L rather than H, $\phi(s_L|\theta = L) > \phi(s_L|\theta = H)$. It can better distinguish the sender's type if it generates s_H more frequently when the sender is high type than when the sender is low type (higher $\phi(s_H|\theta = H)/\phi(s_H|\theta = L)$). Similarly, it can better distinguish the sender's type if it generates s_L more frequently when the sender is low type than when the sender is high type (higher $\phi(s_L|\theta = L)/\phi(s_L|\theta = H)$). This motivates the following definition, which is useful in the subsequent analyses.

Definition 2. A classifier has a high capacity if $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) \geq (1-\rho)\Delta_L^R/(\rho\Delta_H^R)$ and $\phi(s_L|\theta = L)/\phi(s_L|\theta = H) \geq (\Delta_L^S - C)\rho\Delta_H^R/[\Delta_L^S\rho\Delta_H^R - (1-\rho)\Delta_L^R C]$. Otherwise, it has a low capacity.

Both thresholds increase with Δ_L^R and decrease with Δ_H^R. Intuitively, a higher Δ_L^R implies that the receiver incurs a greater loss from taking the wrong action, thereby requiring a more informative detector to support decision-making. So, the threshold for high capacity is higher. In contrast, a higher Δ_H^R indicates that the receiver benefits more from taking the right action (or equivalently, suffers less from taking the wrong action), reducing the need for a highly informative detector. Hence, the threshold for high capacity is lower. Finally, a higher ρ means that the sender is more likely to be high-type, thus lowering the threshold.

Maximizing Receiver's Payoff

In this case, the designer's problem is: $\max_\beta \mathbb{E}U^R(\beta,\alpha^*(\beta;\phi))$.

Proposition 3. *If the classifier has a low capacity, the optimal true-positive rate of the detector that maximizes the receiver's expected payoff is any* $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$, *which minimizes the low-type sender's equilibrium probability of lying. If the classifier has a high capacity, the optimal true-positive rate*

is any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ *if the lying cost is high,* $C \geq \hat{C}$*, and is* $\beta = \phi(s_L|\theta = L)$ *if the lying cost is low,* $C < \hat{C}$*.*[16]

The receiver benefits from making a more informed decision and wants to better distinguish between the two types of senders. The designer can affect equilibrium outcomes through two channels - disinformation deterrence and information provision. Different detectors induce different equilibrium probabilities of lying because lying is a strategic decision by low-type senders. A detector can benefit the receiver by reducing the percentage of disinformation in equilibrium. For a given amount of disinformation, a detector can also help the receiver by providing more informative alarm signals.

When the lying cost is high, it is easy for the detector to discourage low-type senders from lying. When the classifier has a low capacity, it is hard to provide highly informative alarm signals. In both cases, it is optimal for the designer to rely on the first channel, disinformation deterrence, to maximize the receiver's payoff. The designer minimizes the equilibrium probability of lying by taking advantage of its discrete downward jump at $\hat{\beta}$. When the lying cost is low and the classifier has a high capacity, it is hard to discourage low-type senders from lying but easy to provide highly informative alarm signals. So, the designer uses the second channel, information provision, to maximize the receiver's payoff by taking full advantage of the region $(0 < \beta < \phi(s_L|\theta = L))$ where a unit increase in the true-positive rate corresponds to a small increase in the false-positive rate.

The optimal true-positive rates lie within an intermediate range for the following reasons. A low-type sender has a strong incentive to lie if the detector's true positive rate is low. The receiver will take action r_L and obtain zero payoff upon observing an alarm because the sender is highly likely to be type L. Even in the absence of an alarm, the combination of a high probability of lying and a low detection rate implies that the receiver still has much uncertainty about the sender's type (a weak detector generates a small persuasive effect). Therefore, the receiver either follows a mixed strategy and obtains zero payoff or takes action r_H but earns a low payoff because of the high chance of taking the wrong action when the sender's true type is L. So, a low true-positive rate is not optimal for the receiver.

Upon observing an alarm, the receiver will be fairly confident that the sender is a low type if the detector's true positive rate is high because of the large dissuasive effect of a strong detector. In the meantime,

[16]The threshold $\hat{C} := [\phi(s_H|\theta = H) - (1-\rho)\Delta_L^R\phi(s_H \mid \theta = L)/\rho\Delta_H^R]\phi(s_H \mid \theta = L)\Delta_L^S/\{[\phi(s_H|\theta = H) - (1-\rho)\Delta_L^R\phi(s_H \mid \theta = L)/\rho\Delta_H^R - 1]\phi(s_H \mid \theta = L) + \phi(s_H \mid \theta = H)\}$. Note that among the space of exogenous detectors $\{(\beta, \alpha)|0 < \alpha < \beta\}$, multiple equilibria exist in a set of measure zero. However, when we study the endogenous detector design, different detectors may lead to the same equilibrium payoffs. So, the optimal detector may be a set of detectors.

the receiver will have a high posterior belief about the sender's type and will always take action r_H after receiving no alarm (a strong detector generates a large persuasive effect). In $\rho(1-\alpha)$ amount of time, the sender is a high type, and the receiver earns a positive payoff of Δ_H^R. In $(1-\rho)\sigma^S(1-\beta)$ amount of time, the sender is a low type, and the receiver earns a negative payoff of $-\Delta_L^R$. As the true-positive rate β in increases, the false-positive rate α also increases. So, both the benefit and the cost of action r_H are reduced. According to Lemma 5 and Figure 9, the detector's false-positive rate increases faster in its true-positive rate when the true-positive rate is high. In such cases, the benefit of action r_H decreases at a high rate as β increases. So, a high true-positive rate is also not optimal for the receiver.

Maximizing High-type Sender's Payoff

In this case, the designer's problem is: $\max_\beta \mathbb{E}U_H^S(\beta, \alpha^*(\beta;\phi))$.

Proposition 4. *If the classifier has a low capacity, the optimal true-positive rate of the detector that maximizes the high-type sender's expected payoff is any* $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta=L)\}]$. *If the classifier has a high capacity, the optimal true-positive rate is* $\beta_1 := [\rho\Delta_H^R - (1-\rho)\Delta_L^R]/[\rho\Delta_H^R\phi(s_L|\theta=H)/\phi(s_L|\theta=L) - (1-\rho)\Delta_L^R]$, *which is lower than* $\hat{\beta}$ *and decreases in* $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)$.

A high-type sender wants the receiver to take action r_H as frequently as possible. In particular, he prefers the receiver to always take action r_H given message m_H and no alarm. The minimum true-positive rate required to induce such behavior is lower under a high-capacity classifier than under a low-capacity classifier. This is because, for a given true-positive rate, a higher-capacity classifier yields a lower false-positive rate. This leads to a larger persuasive effect, making it easier to induce the receiver to take action r_H. So, given a higher-capacity classifier, the detector can induce the desired outcome with a lower β. The high-type sender does not want to further increase β because a higher true-positive rate corresponds to a higher false-positive rate. As a result, the detector is more likely to send a false alarm when β is higher. Because the receiver takes action r_L with a positive probability when there is an alarm, the more frequent false alarm hurts the sender's payoff.

The classifier is better at distinguishing between two types of senders if $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)$ is higher. When the classifier has a high capacity, *counter-intuitively*, the optimal true-positive rate decreases in the classifier's capacity - the optimal detector alarms a smaller percentage of disinformation when its underlying classifier is better at distinguishing the sender's type. The mechanism is the following. A high-type sender wants to choose the lowest true-positive rate that induces the receiver to always take the sender's

desired action r_H upon seeing message m_H and no alarm. Fixing a true-positive rate, the detector has a lower false-positive rate if the classifier has a higher capacity. So, the receiver's posterior belief after observing message m_H and no alarm is higher. The detector can induce the receiver to take action r_H even if its true-positive rate is adjusted downward. This way, the sender can still obtain the highest payoff and is less likely to be the object of a false alarm.

Maximizing a Weighted Sum of the Sender's and Receiver's Payoffs

The designer may care about both the sender's and the receiver's payoffs when choosing the detector. In addition to the commonly studied objective of maximizing social welfare, the designer may have more complex incentives. For example, a newly launched e-commerce platform may care about both sellers (senders) and consumers (receivers), but focuses on enhancing the consumer experience to grow market share, whereas other platforms may prioritize retaining high-quality sellers. To capture a broad range of the designer's objectives, we consider an objective function consisting of a weighted sum of the expected payoffs of the receiver, the high-type sender, and the low-type sender, $\max_\beta \mathbb{E}\widetilde{W}(\beta) := \mathbb{E}U^R(\beta, \alpha^*(\beta;\phi)) + w_H\rho\mathbb{E}U_H^S(\beta, \alpha^*(\beta;\phi)) + w_L(1-\rho)\mathbb{E}U_L^S(\beta, \alpha^*(\beta;\phi))$, where $w_R, w_H > 0$. In the special case where both weights w_H and w_L are one, the designer's objective becomes maximizing social welfare $\max_\beta \mathbb{E}W(\beta) := \mathbb{E}U^R(\beta, \alpha^*(\beta;\phi)) + \rho\mathbb{E}U_H^S(\beta, \alpha^*(\beta;\phi)) + (1-\rho)\mathbb{E}U_L^S(\beta, \alpha^*(\beta;\phi))$.

Proposition 5. *If the classifier has a low capacity, any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ is optimal. If the classifier has a high capacity, the set of optimal true-positive rates is:*

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < \tilde{C}(w_H, w_L) \\ \mathcal{B}(w_H, w_L) \cup [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}] & \text{if } C = \tilde{C}(w_H, w_L) \\ [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}] & \text{if } C \in (\tilde{C}(w_H, w_L), \Delta_L^S(1-\beta_1)] \end{cases},$$

where

$$\mathcal{B}(w_H, w_L) := \begin{cases} \{\phi(s_L \mid \theta = L)\} & \text{if } n_0 w_H + l_0 w_L < 1 \\ [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } n_0 w_H + l_0 w_L = 1 \\ \{\beta_1\} & \text{if } n_0 w_H + l_0 w_L > 1 \end{cases},$$

$\tilde{C}(w_H, w_L)$ is a continuous function increasing in w_H and w_L; n_0 and l_0 are strictly positive constants.[17]

[17] The specific expressions for $\tilde{C}(w_H, w_L)$, n_0, and l_0 are in the online appendix.

If the designer has access to a classifier with a low capacity, the receiver's and both senders' preferences towards the detector are aligned.[18] So, the optimal detector for each individual also maximizes the weighted sum of the sender's and receiver's payoffs. If the designer has access to a classifier with a high capacity, however, their preferences towards the detector are not aligned. The senders prefer a lower true-positive rate than the receiver. When the designer cares about both senders and receivers, the optimal detector reflects a compromise between their preferences. As the designer places greater weight on senders (higher w_H and w_L), the optimal true-positive rate decreases. This is because senders prefer a lower β than receivers, and the optimal detector adjusts accordingly to balance these competing interests.

Comparison With the No False-positive Alarm Benchmark

Compared to the no false-positive benchmark in section 3.2, the consideration of false positives leads to qualitatively different results of detector design regardless of the designer's objective. When there are no false-positive alarms, there is no trade-off in the detector design, and the detector designer always prefers a higher true-positive rate. In contrast, we have shown that the designer strictly prefers an intermediate true-positive rate to the highest true-positive rate in the presence of false-negative alarms. A higher true-positive rate may reduce the receiver's expected payoff, the high-type sender's expected payoff, and social welfare when we take into account the possibility of false-positive alarms and the players' strategic responses.

5 Extensions

5.1 Restriction on the Alarm Rule: $\lambda_H = 0$

In this section, we consider the case where the detector never sends an alarm when the classifier predicts signal s_H. In this case, the space of the feasible detectors given a classifier ϕ becomes $\{(\beta, \alpha) | \beta = \phi(s_L|\theta = L)\lambda_L, \alpha = \phi(s_L|\theta = H)\lambda_L, \lambda_L \in [0, 1]\}$. The following result characterizes the optimal true-positive rate.

Proposition 6. *1. (Maximizing receiver's payoff) If the classifier has a low capacity, the optimal true-positive rate is any $\beta \in [\hat{\beta}, \phi(s_L|\theta = L)]$ when $C \geq \Delta_L^S \phi(s_H \mid \theta = L)$, which minimizes the low-type sender's equilibrium probability of lying, and is any $\beta \in [0, \phi(s_L|\theta = L)]$ when $C < \Delta_L^S \phi(s_H \mid \theta = L)$. If the classifier has a high capacity, the optimal true-positive rate is any $\beta \in [\hat{\beta}, \phi(s_L|\theta = L)]$ when $C \geq \Delta_L^S \phi(s_H \mid \theta = L)$, is $\phi(s_L|\theta = L)$ when $C < \Delta_L^S \phi(s_H \mid \theta = L)$*

[18] We characterize the optimal true-positive rate for the low-type sender in the online appendix A.9.

and $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) > (1-\rho)\Delta_L^R/(\rho\Delta_H^R)$*, and is any* $\beta \in [0, \phi(s_L|\theta = L)]$ *when* $C < \Delta_L^S \phi(s_H \mid \theta = L)$ *and* $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) = (1-\rho)\Delta_L^R/(\rho\Delta_H^R)$.

2. *(Maximizing high-type sender's payoff) If the classifier has a low capacity, the optimal true-positive rate is any* $\beta \in [\hat{\beta}, \phi(s_L|\theta = L)]$ *when* $C \geq \Delta_L^S \phi(s_H \mid \theta = L)$*, and is* $\phi(s_L|\theta = L)$ *when* $C < \Delta_L^S \phi(s_H \mid \theta = L)$*. If the classifier has a high capacity, the optimal true-positive rate is* β_1.

3. *(Maximizing a weighted sum of sender's and receiver's payoffs* $\mathbb{E}\widetilde{W}(\beta)$*) If the classifier has a low capacity, the optimal true-positive rate is any* $\beta \in [\hat{\beta}, \phi(s_L|\theta = L)]$ *when* $C \geq \Delta_L^S \phi(s_H \mid \theta = L)$*, and is* $\phi(s_L|\theta = L)$ *when* $C < \Delta_L^S \phi(s_H \mid \theta = L)$*. If the classifier has a high capacity, the optimal true-positive rate is:*

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < \max\left\{\Delta_L^S \phi(s_H \mid \theta = L), (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S\right\} \\ \{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S > \Delta_L^S \phi(s_H \mid \theta = L) \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{otherwise} \end{cases}.$$

When the lying cost is high, the detector can significantly reduce the incentives for low-type senders to lie, even if it sends an alarm at a low frequency ($\lambda_H = 0$). In this case, the low-type sender's equilibrium lying probability lies within the right part of Figure 5 and stays at a low level. The results and underlying mechanisms are similar to the main model. When the lying cost is low, low-type senders have a strong incentive to lie. To reduce the probability of lying, the designer prefers a detector with a high true-positive rate, which may require sending out alarms with a positive probability even if the classifier predicts s_H ($\lambda_H > 0$). When such a strategy is not feasible in this extension, the low-type sender's equilibrium lying probability lies within the left part of Figure 5 and remains at a high level because of the restriction on the alarm rule.

5.2 Endogenous Commission Fee and Strategic Platform Entry

The main model has considered various objective functions of the detector designer. While these objectives are relevant to some practical applications, real-world designers may act more strategically. The implications of such strategic behavior depend on the institutional details of the problems. In this extension, we analyze a particular setting in which an e-commerce platform (designer) chooses the price (i.e., com-

mission fee) in addition to the information structure of the detector. Platform pricing and detector design jointly affect the seller's (sender's) platform entry decision, which in turn affects the designer's payoff. In particular, we consider a setting in which a platform charges an endogenous commission rate f to sellers and designs a disinformation detector, keeping the assumption from the previous section that the platform does not raise an alarm when the classifier predicts s_H. We consider an exogenous product price P for two reasons. First, it allows us to focus on the strategic decision of the detector designer. Second, because sellers have private information about their types, considering endogenous product price will introduce another dimension of signaling (seller's message and price), which significantly complicates the problem.

Sellers choose whether to enter the platform. Consistent with the main model, ρ percentage of sellers are high-quality (high type), whereas the remaining $1-\rho$ percentage are low-quality (low type). Low-quality sellers have a reservation utility of $u_L > 0$, and high-quality sellers have a reservation utility of $u_H > 0$. The positive reservation utility makes sellers' entry decisions non-trivial. It captures the idea that sellers can make profits through other channels. After entry, low-quality sellers can choose to misrepresent their quality at a cost of $C > 0$, which may trigger the platform's alarm. The platform endogenously chooses the commission rate f and the detector $\{\beta, \alpha\}$ to influence the entry decisions of sellers. The platform's goal is to maximize its total commission fee from transactions. To simplify the problem, we keep the restriction on the alarm rule in the previous section, $\lambda_H = 0$.

We normalize the consumer's valuation for a high-quality product to 1 and denote the valuation for a low-quality product by v. We assume that the consumer gains a positive payoff of $1-P$ from purchasing a high-quality product and a negative payoff of $v-P$ from purchasing a low-quality product, $v < P < 1$. Consistent with the main model, we assume that the consumer does not purchase a product without additional information about the sender's type other than the prior, $\rho(1-P)+(1-\rho)(v-P) < 0$.

The game proceeds as follows: First, the platform decides the commission rate and the detector design. Second, sellers simultaneously decide whether to enter the platform. Third, the entry decisions are publicly revealed, and each seller chooses the messages. Fourth, the detector sends a signal for each message. Lastly, consumers decide whether to purchase a product from each seller. The following proposition characterizes the optimal commission rate and detector.

Proposition 7. *1. If the lying cost is low, $C < \rho(1-P)P[1-(1-u_H/P)(P-v)/(1-v)]/[(1-\rho)(P-v)] - u_L$, and the classifier's capacity is sufficiently high, $\phi(s_L|\theta=L)/\phi(s_L|\theta=H) \geq a$*

constant $\bar{\phi}$ *and* $\phi(s_H|\theta=H)/\phi(s_H|\theta=L) \geq (1-\rho)(P-v)/[\rho(1-P)]$*, the optimal detector and commission fee are*

$$\beta^* = \eta, \alpha^* = \alpha^*(\eta,\phi) = \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\eta, \text{ and } f^* = 1 - \max\left\{\frac{u_H}{1-\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\eta}, \frac{C+u_L}{1-\eta}\right\}\frac{1}{P},$$

$$\text{where } \eta = \left[\frac{(1-\rho)(P-v)}{\rho(1-P)} - 1\right] / \left[\frac{(1-\rho)(P-v)}{\rho(1-P)} - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\right].$$

Both types of sellers enter the market.

2. *In other cases, the optimal commission fee is* $f^* = 1 - u_H/P$*. Only high-quality sellers enter the market given any detector.*

When both types of sellers enter the market, the properties of the optimal detector remain consistent with those in the main model. In this case, platform pricing (i.e., the choice of the commission fee) acts as a *strategic complement* to detector design, providing an additional lever for the platform to extract surplus from sellers. In contrast, the detector becomes irrelevant when the commission rate is sufficiently high to ensure that only high-quality sellers participate, because consumers will then always purchase the product. In this case, platform pricing serves as a *strategic substitute* for detector design, deterring low-quality sellers from entering.

The platform is better off inducing entry by both seller types rather than retaining only high-quality sellers when the lying cost is low and the classifier's capacity is sufficiently high. The intuition is as follows. The platform can charge a high commission rate if it limits participation to high-quality sellers. By lowering the commission rate, it can encourage broader seller participation, which brings both benefits and costs. It gains additional commission revenue from low-quality sellers but sacrifices some profits from high-quality sellers. When the lying cost is lower or the classifier is more effective, the platform can attract low-quality sellers with a smaller discount on the commission rate, thereby reducing the cost and increasing the benefit of inducing both types of sellers to enter the market.

6 Discussion and Concluding Remarks

Disinformation detection is becoming increasingly important and relevant because it is easier than ever to create and disseminate disinformation. To study the strategic interaction between disinformation generation

and detection, this paper considers a game-theoretic model where a sender strategically communicates his type to a receiver, and a lie detector generates a noisy signal on the truthfulness of the sender's message. The receiver then infers the sender's type through messages from the sender and the detector.

Because of practical limitations, the detector may make two types of mistakes. It may fail to send an alarm when the sender is lying (false negative). It may also send a false alarm when the sender is truthful (false positive). Previous work has focused on the first type of mistake by implicitly assuming that the false-positive rate is zero. In reality, the sender cannot avoid making the second type of mistake unless he never sends an alarm. Type I (false positive) and type II (false negative) errors may be viewed conceptually as similar to one another in that both types of errors make information less precise. Nevertheless, this paper shows that they generate different effects on strategic communication, which in turn provides important implications for designing the disinformation detector.

We first study how the detection technology affects the equilibrium outcomes. We find a non-monotonic relationship between the sender's probability of lying and the detection accuracy. A stronger detector increases the sender's probability of lying when the true-positive rate is low, because of a persuasive effect, whereas a stronger detector decreases the sender's probability of lying when the true-positive rate is high, because of a dissuasive effect.

We then characterize the optimal detector design under various objectives. The designer always chooses the lowest feasible false-positive rate given any true-positive rate. The possibility of false-positive alarms implies that the designer will not choose the largest true-positive rate. Instead, the designer chooses different intermediate true-positive rates for different objectives. Counter-intuitively, the optimal detector may raise an alarm about a smaller percentage of disinformation when its underlying classifier is better at distinguishing the sender's type.

Our results have important managerial implications. Regarding the descriptive value, we find qualitatively different insights about the relationship between the sender's probability of lying and the detector's accuracy when we allow for false-positive alarms. Without false-positive alarms, an alarm eliminates all the uncertainty about the sender's type. The receiver's posterior belief goes all the way to zero after observing an alarm. Thus, two detectors with different true-positive rates generate the same *dissuasive effect*. In contrast, in the presence of false-positive alarms, two detectors with the same false-positive rate but different true-positive rates generate different dissuasive effects. Variations in the dissuasive effects lead to the non-monotonic relationship between the sender's probability of lying and the detector's accuracy.

Regarding the prescriptive value, we characterize the optimal design of the detector in the presence of practical limitations. Importantly, the possibility of false-positive alarms implies that the designer should not choose the largest true-positive rate. Instead, the designer should choose different intermediate true-positive rates given different objectives. The optimal detector may even raise alarms about a smaller percentage of disinformation when its underlying classifier is better at distinguishing the sender's type. The qualitatively different and counter-intuitive findings highlight the importance of considering the interaction between senders' strategic behavior and both types of mistakes by the detection technology in practice.

The model can be extended to incorporate additional real-world features. In the baseline setup, we assume the detector generates binary signals: an alarm if it suspects the message is disinformation, and no alarm if it deems the message trustworthy. Both signals are noisy because of the potential for false positives and false negatives. In practice, however, the platform or regulator may sometimes be able to verify the validity of a message. For instance, if a reviewer on Amazon claims to have purchased a product twice recently because of its high quality, Amazon can verify this claim by examining the user account's transaction history. In such cases, the detector may produce three possible outcomes: no alarm, alarm, and verification of the message content. The first two outcomes remain noisy signals, as in the main model, whereas the third fully resolves uncertainty about the sender's message. When the probability of verifying the content is very high, the likelihood that a low-type sender can successfully mimic a high-type sender is very low, meaning the benefit of lying is outweighed by its cost. In such cases, senders never lie in equilibrium. In other cases, while the incentive to lie is reduced, a low-type sender will still lie with positive probability. The resulting equilibrium remains qualitatively similar to that of the main model.

We follow the standard assumption of a homogeneous lying cost in the costly lying literature. It is, however, plausible that senders differ in their lying costs, with some facing high costs and others facing low costs. In such cases, equilibrium behavior varies with the strength of the detector. When the detector is strong, low-cost senders will mix between lying and truth-telling, whereas high-cost senders will never lie. Conversely, when the detector is weak, low-cost senders will always lie, and high-cost senders will adopt a mixed strategy. As discussed in Section 4, mixed strategies play a central role in the model's mechanisms. Because only one type of sender (either high-cost or low-cost) adopts mixed strategies in equilibrium, introducing heterogeneity in lying costs does not qualitatively alter the equilibrium structure or the core mechanisms of the model.

There are other interesting areas for future research. In this paper, the sender cannot affect the detector's

ability. Future research can consider the possibility that a sender can make an effort to affect the detector's ability. It would also be interesting to extend the sender's type from binary to a continuous type, which may generate additional insights. Lastly, we study the optimal detector design for a given classifier. This setup reflects the fact that it is much harder to change the capacity of the classifier than to change how to use an endowed classifier to detect disinformation and send alarms because it takes lots of time, money, and data to train the classifier. Nevertheless, it may be interesting to also endogenize the capacity of the classifier when it is feasible to change the classifier in some applications.

Appendix A Proof

A.1 Proof of Lemma 3 and Proposition 1

For completeness, we analyze all cases including $0 < \alpha = \beta$. We divide the analysis into two cases.

Case 1: Complete Lying ($\sigma^S = 1$) When the low-type sender always sends message m_H, m_L becomes an off-path message. In this case, the receiver's belief and action after receiving m_L can be arbitrary in a PBE, as long as the sender has no profitable deviation to m_L.

Lemma 7. *A PBE with $\sigma^S = 1$ exists if and only if $\alpha \leq 1 - (1-\rho)\Delta_L^R(1-\beta)/(\rho\Delta_H^R)$ and $\beta \leq \hat{\beta} := 1 - C/\Delta_L^S$. If $\alpha < 1 - (1-\rho)\Delta_L^R(1-\beta)/(\rho\Delta_H^R)$, the set of equilibria is $\{(\sigma_a^R = 0, \sigma_{na}^R = 1, \sigma_{L,na}^R, \sigma^S = 1) : \sigma_{L,na}^R \leq 1 - \beta - C/\Delta_L^S\}$. If $\alpha = 1 - (1-\rho)\Delta_L^R(1-\beta)/(\rho\Delta_H^R)$ and $\beta \leq \hat{\beta}$, the set of equilibria is $\{(\sigma_a^R = 0, \sigma_{na}^R, \sigma_{L,na}^R, \sigma^S = 1) : \sigma_{L,na}^R \leq (1-\beta)\sigma_{na}^R - C/\Delta_L^S, \sigma_{na}^R \in [C/[(1-\beta)\Delta_L^S], 1]\}$.*

Proof. If $\alpha = \beta$ and the low-type sender always lies $\sigma^S = 1$, then the receiver always takes action $r = r_L$ because she does not get new information about the sender's type other than the prior. But then, the low-type sender can be better off by not lying. There is no such PBE. So, we only need to consider $\beta > \alpha$.

In a PBE where m_L is an off-path message, both types of sender send m_H. Avoiding profitable deviation of low-type sender requires $\sigma_{L,na}^R \leq \beta\sigma_a^R + (1-\beta)\sigma_{na}^R - C/\Delta_L^S$ and avoiding profitable deviation of high-type sender requires $\sigma_{L,na}^R - C/\Delta_H^S \leq \alpha\sigma_a^R + (1-\alpha)\sigma_{na}^R$. By Lemma 8, $\beta\sigma_a^R + (1-\beta)\sigma_{na}^R \leq \alpha\sigma_a^R + (1-\alpha)\sigma_{na}^R$, so we only requires $\sigma_{L,na}^R \leq \beta\sigma_a^R + (1-\beta)\sigma_{na}^R - C/\Delta_L^S < 1$. By the sequential rationality, the receiver's belief off the equilibrium path, $b(H \mid m_L, na)$, can be any value less than or equal to $\hat{\rho}$.

With $\sigma^S = 1$, the receiver's on-path beliefs are $b(H \mid m_H, a) = \alpha\rho/[\alpha\rho + \beta(1-\rho)]$ and $b(H \mid m_H, na) = (1-\alpha)\rho/[(1-\alpha)\rho + (1-\beta)(1-\rho)]$. As $\beta > \alpha$, we have $b(H \mid m_H, a) < \rho < \hat{\rho}$. Thus, sequential rationality of the receiver indicates that $\sigma_a^R = 0$. The equilibrium strategies are:

$$\left\{(\sigma_a^R = 0, \sigma_{na}^R, \sigma_{L,na}^R, \sigma^S = 1) : \sigma_{L,na}^R \leq (1-\beta)\sigma_{na}^R - \frac{C}{\Delta_L^S}, \sigma_{na}^R \in \left[\frac{C}{(1-\beta)\Delta_L^S}, 1\right]\right\} \tag{1}$$

Given $\sigma_{na}^R \in [C/[(1-\beta)\Delta_L^S], 1]$, by sequential rationality of the receiver given $\{m_H, na\}$, we have $b(H \mid$

$m_H, na) = (1-\alpha)\rho/[(1-\alpha)\rho + (1-\beta)(1-\rho)] \geq \hat{\rho} := \Delta_L^R/(\Delta_L^R + \Delta_H^R)$, i.e., $\alpha \leq 1 - [(1-\rho)\Delta_L^R/(\rho\Delta_H^R)](1-\beta)$. There exists equilibrium fulfill the conditions in (1) if and only if $\exists \sigma_{L,na}^R \leq (1-\beta)\sigma_{na}^R - C/\Delta_L^S$ for some $\sigma_{na}^R \in \left[C/[(1-\beta)\Delta_L^S], 1\right]$, which is given by $\beta \leq \hat{\beta} := 1 - C/\Delta_L^S$. Therefore, a PBE with $\sigma^S = 1$ exists if and only if $\alpha \leq 1 - [(1-\rho)\Delta_L^R/(\rho\Delta_H^R)](1-\beta)$ and $\beta \leq \hat{\beta}$, which is given by (1) with belief $b(H \mid m_L, na) \leq \hat{\rho}$, $b(H \mid m_H, a) = \alpha\rho/[\alpha\rho + \beta(1-\rho)]$, and $b(H \mid m_H, na) = (1-\alpha)\rho/[(1-\alpha)\rho + (1-\beta)(1-\rho)]$. □

Case 2: Partial Lying ($\sigma^S < 1$) In a Perfect Bayesian Equilibrium (PBE) where $m = m_L$ is an on-path message (i.e., $\sigma^S < 1$), the PBE definition requires the receiver to maintain a consistent belief, specifically $b(L \mid m_L, na) = 1$. This implies that any sender who chooses $m = m_L$ must be of type $\theta = L$. Consequently, the receiver always selects action $r = r_L$ upon receiving $m = m_L$ (i.e., $\sigma_{L,na}^R = 0$).

For a sender with type $\theta = L$, the expected payoff under strategy σ^S is given by $\mathbb{E}U_0^S(\sigma^S; \{\sigma_{na}^R, \sigma_a^R\}) = \sigma^S[\left(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R\right)\Delta_L^S - C]$. The sender's best response to $\{\sigma_{na}^R, \sigma_a^R\}$ is characterized by:

$$\sigma_{\text{BR}}^S(\{\sigma_{na}^R, \sigma_a^R\}) \begin{cases} = 1, & \left(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R\right)\Delta_L^S > C \\ \in [0,1], & \left(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R\right)\Delta_L^S = C \\ = 0, & \left(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R\right)\Delta_L^S < C \end{cases} \tag{BR1}$$

The receiver's best response, maintaining consistent beliefs, to σ^S is given by:

$$\sigma_{a,\text{BR}}^R(\sigma^S) \begin{cases} = 1, & \sigma^S < \frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R} \\ \in [0,1], & \sigma^S = \frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R} \\ = 0, & \sigma^S > \frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R} \end{cases}; \quad \sigma_{na,\text{BR}}^R(\sigma^S) \begin{cases} = 1, & \sigma^S < \frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R} \\ \in [0,1], & \sigma^S = \frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R} \\ = 0, & \sigma^S > \frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R} \end{cases} \tag{BR2}$$

We now characterize the equilibrium for a given detector.

Equilibrium Characterization when $\alpha = \beta \in (0,1]$ The receiver's best response simplifies to:

$$\begin{cases} \sigma_{a,\text{BR}}^R(\sigma^S) = \sigma_{na,\text{BR}}^R(\sigma^S) = 1, & \sigma^S < \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \sigma_{a,\text{BR}}^R(\sigma^S) \in [0,1], \sigma_{na,\text{BR}}^R(\sigma^S) \in [0,1], & \sigma^S = \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \sigma_{a,\text{BR}}^R(\sigma^S) = \sigma_{na,\text{BR}}^R(\sigma^S) = 0, & \sigma^S > \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \end{cases}$$

Consider first the case where $\sigma^S < \rho\Delta_H^R/[(1-\rho)\Delta_L^R]$. The receiver's best response implies $\sigma_a^R = \sigma_{na}^R = 1$ in equilibrium. Because $\sigma^S \in [0,1)$, the equilibrium condition requires $\Delta_L^S = \left(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R\right)\Delta_L^S \leq C$ (with equality if $\sigma^S > 0$). However, given that $C < \min\{\Delta_H^S, \Delta_L^S\}$, this leads to a contradiction. Therefore, no equilibrium exists in this case.

Next, consider $\sigma^S = \rho\Delta_H^R/[(1-\rho)\Delta_L^R]$. The equilibrium conditions from both players' best responses require that σ_a^R and σ_{na}^R satisfy $\left(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R\right)\Delta_L^S = C$. There exists an equilibrium $\sigma^S = \rho\Delta_H^R/[(1-\rho)\Delta_L^R]$, σ_{na}^R, σ_a^R where $\left(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R\right)\Delta_L^S = C$.

Finally, consider $\sigma^S > \rho\Delta_H^R/[(1-\rho)\Delta_L^R]$. The receiver's best response yields $\sigma_a^R = \sigma_{na}^R = 0$ in

equilibrium. Because $\sigma^S > 0$, the equilibrium condition requires $0 = \left(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R\right)\Delta_L^S \geq C$, which is impossible. Therefore, no equilibrium exists in this case either.

Equilibrium when $0 < \alpha < \beta$ (Proposition 1)

1. Separating equilibrium with $\sigma^S = 0$ implies $\sigma_{na}^R = \sigma_a^R = 1$, which requires $\Delta_L^S \leq C$. There is a contradiction to the definition of C, so there is no complete separating equilibrium.

2. Semi-separating equilibrium with $\sigma^S \in (0,1)$ requires $(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R)\Delta_L^S = C$. By the Lemma 8 and (BR2), we can discuss potential equilibria by following cases:

   (a) $\sigma_{na}^R \in (0,1)$ and $\sigma_a^R = 0$: this case requires $\sigma_{na}^R = C/[(1-\beta)\Delta_L^S] \in (0,1)$ (i.e., $\beta < \hat{\beta}$) and $\sigma^S = [(1-\alpha)\rho\Delta_H^R]/[(1-\beta)(1-\rho)\Delta_L^R] < 1$ (i.e., $\alpha > 1 - [(1-\rho)\Delta_L^R/(\rho\Delta_H^R)](1-\beta)$).

   (b) $\sigma_{na}^R = 1$ and $\sigma_a^R = 0$: this case requires $\sigma_{na}^R = C/[(1-\beta)\Delta_L^S] = 1$ (i.e., $\beta = \hat{\beta}$) and $\sigma^S \in [(\alpha\rho\Delta_H^R)/[\beta(1-\rho)\Delta_L^R], \min\{[(1-\alpha)\rho\Delta_H^R]/[(1-\beta)(1-\rho)\Delta_L^R], 1\}]$.

   (c) $\sigma_{na}^R = 1$ and $\sigma_a^R \in (0,1)$: this case requires $\sigma_a^R = C/(\beta\Delta_L^S) - (1-\beta)/\beta \in (0,1)$ (i.e., $\beta > \hat{\beta}$) and $\sigma^S = (\alpha\rho\Delta_H^R)/[\beta(1-\rho)\Delta_L^R]$.

Equilibrium when $0 = \alpha < \beta$ (Lemma 3) By Lemma 7, the PBEs when $\alpha = 0$ are the following:

Separating equilibrium $\sigma^S = 0$: It implies that $\sigma_{na}^R = 1$ and the (m_H, a) is off the equilibrium path. The existence of a PBE requires $(\beta\sigma_a^R + 1 - \beta)\Delta_L^S \leq C$, i.e., $\sigma_a^R \leq C/(\beta\Delta_L^S) - (1-\beta)/\beta$. There exists such $\sigma_a^R \in [0,1]$ if and only if $C/(\beta\Delta_L^S) - (1-\beta)/\beta \geq 0$, which is given by $\beta \geq \hat{\beta}$. The corresponding belief requires $b(H \mid m_H, a) \leq \hat{\rho}$.

Semi-separating equilibrium $\sigma^S \in (0,1)$: It implies that $(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R)\Delta_L^S = C$. By Lemma 8 and (BR2), we can discuss potential equilibria by following cases:

1. $\sigma_{na}^R \in (0,1)$ and $\sigma_a^R = 0$: this case requires $\sigma_{na}^R = C/[(1-\beta)\Delta_L^S] \in (0,1)$ (i.e., $\beta < \hat{\beta}$) and $\sigma^S = \rho\Delta_H^R/[(1-\beta)(1-\rho)\Delta_L^R] < 1$ (i.e., $\beta < 1 - \rho\Delta_H^R/[(1-\rho)\Delta_L^R]$).

2. $\sigma_{na}^R = 1$ and $\sigma_a^R = 0$: this case requires $\sigma_{na}^R = C/[(1-\beta)\Delta_L^S] = 1$ (i.e., $\beta = \hat{\beta}$) and $\sigma^S \in [0, \min\{\rho\Delta_H^R/[(1-\beta)(1-\rho)\Delta_L^R], 1\}]$.

3. $\sigma_{na}^R = 1$ and $\sigma_a^R \in (0,1)$: this case requires $\sigma_a^R = C/(\beta\Delta_L^S) - (1-\beta)/\beta \in (0,1)$ (i.e., $\beta > \hat{\beta}$) and $\sigma^S = 0$.

1. if $C \geq [(\rho\Delta_H^R)/[(1-\rho)\Delta_L^R]]\Delta_L^S$,

$$\begin{cases} \sigma^{S*} = \frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^{R\,*} = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^{R*} = 0, & 0 < \beta < \hat{\beta} := \hat{\beta} \\ \sigma^{S*} \in \left[0, \frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}\right], \sigma_{na}^{R\,*} = 1, \sigma_a^{R*} = 0, & \beta = \hat{\beta} \\ \sigma^{S*} = 0, \sigma_{na}^{R\,*} = 1, \sigma_a^{R*} \leq \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta > \hat{\beta} \end{cases}$$

2. if $C < [(\rho\Delta_H^R)/[(1-\rho)\Delta_L^R]]\Delta_L^S$,

$$\begin{cases} \sigma^{S*} = \frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^{R\,*} = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^{R*} = 0, & 0 < \beta < 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \sigma^{S*} = 1, \sigma_{na}^{R\,*} = \left[\frac{C}{(1-\beta)\Delta_L^S}, 1\right], \sigma_a^{R*} = 0, & \beta = 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \sigma^{S*} = 1, \sigma_{na}^{R\,*} = 1, \sigma_a^{R*} = 0, & 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} < \beta < \hat{\beta} \\ \sigma^{S*} \in [0,1], \sigma_{na}^{R\,*} = 1, \sigma_a^{R*} = 0, & \beta = \hat{\beta} \\ \sigma^{S*} = 0, \sigma_{na}^{R\,*} = 1, \sigma_a^{R*} \leq \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta > \hat{\beta} \end{cases}$$

A.2 Refinement on Multiple Equilibria

When $\beta = \hat{\beta}$, $\{\alpha\rho\Delta_H^R/[\beta(1-\rho)\Delta_L^R], 1, 0\}$ Pareto dominates other equilibria. When $\beta \in \left(0, \hat{\beta}\right)$ and $\alpha = 1 - (1-\rho)\Delta_L^R(1-\beta)/(\rho\Delta_H^R)$, $\{1, 1, 0\}$ Pareto dominates other equilibria. Table 3 summarizes the equilibrium payoff under this refinement.

β Range \ α Range	$\alpha \leq 1 - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta)$	$\alpha \in \left(1 - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta), \beta\right]$
$\left[\hat{\beta}, 1\right)$	$\mathbb{E}U^R = \left(1 - \frac{\alpha}{\beta}\right)\rho\Delta_H^R$, $\mathbb{E}U_L^S = 0$, $\mathbb{E}U_H^S = \Delta_H^S - \frac{(\Delta_L^S - C)\Delta_H^S}{\Delta_L^S}\frac{\alpha}{\beta}$	
$\left(0, \hat{\beta}\right)$	$\mathbb{E}U^R = (1-\alpha)\rho\Delta_H^R - (1-\beta)(1-\rho)\Delta_L^R$, $\mathbb{E}U_L^S = (1-\beta)\Delta_L^S - C$, $\mathbb{E}U_H^S = (1-\alpha)\Delta_H^S$	$\mathbb{E}U^R = 0$, $\mathbb{E}U_L^S = 0$, $\mathbb{E}U_H^S = \frac{C}{\Delta_L^S}\Delta_H^S\frac{1-\alpha}{1-\beta}$

Table 3: Equilibrium payoff under detector $\{\beta, \alpha\}$ and the Pareto-optimal refinement

If $\beta \in [\hat{\beta}, 1)$, the equilibrium is $\sigma^S = (\alpha\rho\Delta_H^R)/[\beta(1-\rho)\Delta_L^R]$, $\sigma_{na}^R = 1$, $\sigma_a^R = C/(\beta\Delta_L^S) - (1-\beta)/\beta$. Because it is an equilibrium with mixed strategies σ^S and σ_a^R, the low-type sender's expected payoff is 0 and the receiver's expected payoff given an alarm is 0. So, the receiver's expected payoff is $\rho(1-\alpha)\Delta_H^R - (1-\rho)(1-\beta)\sigma^S\Delta_L^R = (1-\alpha/\beta)\rho\Delta_H^R$ and the high-type sender's expected payoff is $(1-\alpha+\alpha\sigma_a^R)\Delta_H^S = \Delta_H^S - [(\Delta_L^S - C)\Delta_H^S/\Delta_L^S](\alpha/\beta)$.

If $\beta \in (0, \hat{\beta})$ and $\alpha \leq 1 - [(1-\rho)\Delta_L^R/(\rho\Delta_H^R)](1-\beta)$, the equilibrium is $\sigma^S = 1$, $\sigma_{na}^R = 1$, $\sigma_a^R = 0$. The receiver's expected payoff is $(1-\alpha)\rho\Delta_H^R - (1-\beta)(1-\rho)\Delta_L^R$, the low-type sender's expected payoff is $(1-\beta)\Delta_L^S - C$, and the high-type sender's expected payoff is $(1-\alpha)\Delta_H^S$.

If $\beta \in (0, \hat{\beta})$ and $\alpha > 1 - [(1-\rho)\Delta_L^R/(\rho\Delta_H^R)](1-\beta)$, the equilibrium is $\sigma^S = [(1-\alpha)\rho\Delta_H^R]/[(1-\beta)(1-\rho)\Delta_L^R]$, $\sigma_{na}^R = C/[(1-\beta)\Delta_L^S]$, $\sigma_a^R = 0$. Because it is an equilibrium with mixed strategies σ^S and σ_{na}^R, the low-type sender's expected payoff is 0 ríg and the receiver's expected payoff given no alarm is 0. As the receiver also gets zero payoff given an alarm, the receiver's expected payoff is 0. And the high-type sender's expected payoff is $(1-\alpha)\sigma_{na}^R\Delta_H^S = (C/\Delta_L^S)\Delta_H^S[(1-\alpha)/(1-\beta)]$.

References

Anderson, E. T. and Simester, D. I. (2014). Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research*, 51(3):249–269.

Balbuzanov, I. (2019). Lies and consequences: The effect of lie detection on communication outcomes. *International Journal of Game Theory*, 48(4):1203–1240.

Becker, G. S. and Stigler, G. J. (1974). Law enforcement, malfeasance, and compensation of enforcers. *The Journal of Legal Studies*, 3(1):1–18.

Berman, R. and Katona, Z. (2013). The role of search engine optimization in search marketing. *Marketing Science*, 32(4):644–651.

Berman, R. and Katona, Z. (2020). Curation algorithms and filter bubbles in social networks. *Marketing Science*, 39(2):296–316.

Berman, R., Zhao, H., and Zhu, Y. (2022). Strategic recommendation algorithms: Overselling and demarketing information designs. *Available at SSRN 4301489*.

Björkegren, D., Blumenstock, J. E., and Knight, S. (2020). Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*.

Branco, F., Sun, M., and Villas-Boas, J. M. (2016). Too much information? information provision and search costs. *Marketing Science*, 35(4):605–618.

Buckland, M. and Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19.

Callander, S. and Wilkie, S. (2007). Lies, damned lies, and political campaigns. *Games and Economic Behavior*, 60(2):262–286.

Calvano, E., Calzolari, G., Denicolo, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297.

Cappelen, A. W., Cappelen, C., and Tungodden, B. (2023). Second-best fairness: The trade-off between false positives and false negatives. *American Economic Review*, 113(9):2458–2485.

Chen, J., Ke, T. T., and Shin, J. (2025). Designing detection algorithms for ai-generated content: Consumer inference, creator incentives, and platform strategy. *Available at SSRN 5271493*.

Chen, L. and Papanastasiou, Y. (2021). Seeding the herd: Pricing and welfare effects of social learning manipulation. *Management Science*, 67(11):6734–6750.

Chen, Y., Du, J., and Lei, Y. (2025). The interactions of customer reviews and price and their dual roles in conveying quality information. *Marketing Science*, 44(1):155–175.

Chen, Y., Huang, J., and Gong, Z. (2024). The strategic failure of emission targets. Unpublished Working Paper.

CMA (2015). Online reviews and endorsements. Available at https://goo.gl/GxZ4J7. Published on February 26.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451.

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management science*, 52(10):1577–1593.

Du, J. and Lei, Y. (2022). Information design of matching platforms when user preferences are bidimensional. *Production and Operations Management*, 31(8):3320–3336.

Dziuda, W. and Salas, C. (2018). Communication with detectable deceit. *Available at SSRN 3234695*.

Eliaz, K. and Spiegler, R. (2019). The model selection curse. *American Economic Review: Insights*, 1(2):127–140.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1):384–394.

Goodin, R. E. (1985). Erring on the side of kindness in social welfare policy. *Policy Sciences*, 18(2):141–156.

Gordon, B. R., Jerath, K., Katona, Z., Narayanan, S., Shin, J., and Wilbur, K. C. (2021). Inefficiencies in digital advertising markets. *Journal of Marketing*, 85(1):7–25.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3):461–483.

Guo, L. (2009). Quality disclosure formats in a distribution channel. *Management Science*, 55(9):1513–1526.

Guo, L. and Zhao, Y. (2009). Voluntary quality disclosure and market interaction. *Marketing Science*, 28(3):488–501.

He, S., Hollenbeck, B., Overgoor, G., Proserpio, D., and Tosyali, A. (2022). Detecting fake-review buyers using network structure: Direct evidence from amazon. *Proceedings of the National Academy of Sciences*, 119(47):e2211932119.

He, S., Hollenbeck, B., and Proserpio, D. (2022). The market for fake reviews. *Marketing Science*, 41(5):896–921.

Iyer, G. and Ke, T. T. (2024). Competitive model selection in algorithmic targeting. *Marketing Science*.

Iyer, G. and Singh, S. (2018). Voluntary product safety certification. *Management Science*, 64(2):695–714.

Iyer, G. and Singh, S. (2022). Persuasion contest: Disclosing own and rival information. *Marketing Science*, 41(4):682–709.

Iyer, G., Yao, Y. J., and Zhong, Z. Z. (2024). Precision-recall tradeoff in competitive targeting. Unpublished Working Paper.

Iyer, G. and Zhong, Z. (2022). Pushing notifications as dynamic information design. *Marketing Science*, 41(1):51–72.

Jerath, K. and Ren, Q. (2021). Consumer rational (in) attention to favorable and unfavorable product information, and firm information design. *Journal of Marketing Research*, 58(2):343–362.

Jin, C., Yang, L., and Hosanagar, K. (2023). To brush or not to brush: Product rankings, consumer search, and fake orders. *Information Systems Research*, 34(2):532–552.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kartik, N. (2009). Strategic communication with lying costs. *The Review of Economic Studies*, 76(4):1359–1395.

Kartik, N., Ottaviani, M., and Squintani, F. (2007). Credulity, lies, and costly talk. *Journal of Economic theory*, 134(1):93–116.

Ke, T. T., Lin, S., and Lu, M. Y. (2022). Information design of online platforms. *HKUST Business School Research Paper*, (2022-070).

Kuksov, D. (2009). Communication strategy in partnership selection. *Quantitative Marketing & Economics*, 7(3).

Kuksov, D. and Lin, Y. (2010). Information provision in a vertically differentiated competitive marketplace. *Marketing Science*, 29(1):122–138.

Lappas, T., Sabnis, G., and Valkanas, G. (2016). The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. *Information Systems Research*, 27(4):940–961.

Lauga, D. O., Ofek, E., and Katona, Z. (2022). When and how should firms differentiate? quality and advertising decisions in a duopoly. *Journal of Marketing Research*, 59(6):1252–1265.

Lee, J.-Y., Shin, J., and Yu, J. (2024). Communicating attribute importance under competition. Unpublished Working Paper.

Liang, A. (2019). Games of incomplete information played by statisticians. *arXiv preprint arXiv:1910.07018*.

Lieberman, M. D. and Cunningham, W. A. (2009). Type i and type ii error concerns in fmri research: re-balancing the scale. *Social cognitive and affective neuroscience*, 4(4):423–428.

Lin, S., Shi, Z. J., and Sun, X. (2025). Towards intelligent shopping assistant: Can llm chatbot empower consumer decision making? *Available at SSRN 5088975*.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management science*, 62(12):3412–3427.

Mattes, K., Popova, V., and Evans, J. R. (2023). Deception detection in politics: Can voters tell when politicians are lying? *Political Behavior*, 45(1):395–418.

Mayzlin, D. (2006). Promotional chat on the internet. *Marketing science*, 25(2):155–163.

Mayzlin, D., Dover, Y., and Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8):2421–2455.

Mayzlin, D. and Shin, J. (2011). Uninformative advertising as an invitation to search. *Marketing science*, 30(4):666–685.

Miklós-Thal, J. and Tucker, C. (2019). Collusion by algorithm: Does better demand prediction facilitate coordination between sellers? *Management Science*, 65(4):1552–1561.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, pages 380–391.

Montiel Olea, J. L., Ortoleva, P., Pai, M. M., and Prat, A. (2022). Competing models. *The Quarterly Journal of Economics*, 137(4):2419–2457.

Ning, Z. E., Shin, J., and Yu, J. (2025). Targeted advertising as implicit recommendation: strategic mistargeting and personal data opt-out. *Marketing Science*, 44(2):390–410.

O'Connor, J. and Wilson, N. E. (2021). Reduced demand uncertainty and the sustainability of collusion: How ai could affect competition. *Information Economics and Policy*, 54:100882.

Papanastasiou, Y. (2020). Fake news propagation and detection: A sequential model. *Management Science*, 66(5):1826–1846.

Pei, A. and Mayzlin, D. (2022). Influencing social media influencers through affiliation. *Marketing Science*, 41(3):593–615.

Piccolo, S., Tedeschi, P., and Ursino, G. (2015). How limiting deceptive practices harms consumers. *The RAND Journal of Economics*, 46(3):611–624.

Piccolo, S., Tedeschi, P., and Ursino, G. (2018). Deceptive advertising with rational buyers. *Management Science*, 64(3):1291–1310.

Qian, K. and Jain, S. (2024). Digital content creation: An analysis of the impact of recommendation systems. *Management Science*.

Qiu, Y. and Rao, R. C. (2024). Can merchants benefit from entry by (amazon-like) platform if multiagent prices signal quality? *Marketing Science*, 43(4):778–796.

Rao, A. and Wang, E. (2017). Demand for "healthy" products: False claims and ftc regulation. *Journal of Marketing Research*, 54(6):968–989.

Rayo, L. and Segal, I. (2010). Optimal information disclosure. *Journal of political Economy*, 118(5):949–987.

Rhodes, A. and Wilson, C. M. (2018). False advertising. *The RAND Journal of Economics*, 49(2):348–369.

Salant, Y. and Cherry, J. (2020). Statistical inference in games. *Econometrica*, 88(4):1725–1752.

Shin, J. (2005). The role of selling costs in signaling price image. *Journal of Marketing Research*, 42(3):302–312.

Shin, J. and Wang, C.-Y. (2024). The role of messenger in advertising content: Bayesian persuasion perspective. *Marketing Science*.

Shulman, J. D. and Gu, Z. (2024). Making inclusive product design a reality: How company culture and research bias impact investment. *Marketing Science*, 43(1):73–91.

Sun, M. (2011). Disclosing multiple product attributes. *Journal of Economics & Management Strategy*, 20(1):195–224.

Sun, M. and Tyagi, R. K. (2020). Product fit uncertainty and information provision in a distribution channel. *Production and Operations Management*, 29(10):2381–2402.

Villas-Boas, J. M. (2004). Communication strategies and product line design. *Marketing Science*, 23(3):304–316.

Wu, Y., Zhang, K., and Xie, J. (2020). Bad greenwashing, good greenwashing: Corporate social responsibility and information transparency. *Management Science*, 66(7):3095–3112.

Yao, Y. (2024). Dynamic persuasion and strategic search. *Management Science*, 70(10):6778–6803.

Zhang, J. (2013). Policy and inference: The case of product labeling. Unpublished Working Paper.

Zheng, X. and Singh, S. (2023). Ambiguous expert communication. *Available at SSRN 4393315*.

Zinman, J. and Zitzewitz, E. (2016). Wintertime for deceptive advertising? *American Economic Journal: Applied Economics*, 8(1):177–192.

Online Appendix for Strategic Disinformation Generation and Detection

A.3 Proof of Lemma 1

We first prove an intuitive result: when there is an alarm, the receiver is less likely to take the sender's desired action r_H compared to when there is no alarm.

Lemma 8. *Given $\beta > \alpha$, $\sigma^R(r_H \mid m_H, a) \leq \sigma^R(r_H \mid m_H, na)$. Specifically, if $\sigma^R(r_H \mid m_H, na) \in [0, 1)$, then $\sigma^R(r_H \mid m_H, a) = 0$. If $\sigma^R(r_H \mid m_H, a) \in (0, 1)$, then $\sigma^R(r_H \mid m_H, na) = 1$.*

Proof. First, we show that $m = m_H$ must be an on-path message in equilibrium. Suppose, for contradiction, that $\sigma^S(m_H \mid H) = \sigma^S(m_H \mid L) = 0$ in a PBE. In this case, the receiver's optimal strategy after receiving $m = m_L$ would be $\sigma^R(r_L \mid m_L, na) = 1$ because $(1-\rho)\Delta_L^R + \rho\Delta_H^R < 0$. However, this creates a profitable deviation for the sender with type $\theta = H$ to send m_H with positive probability. This contradiction proves that $m = m_H$ must be an on-path message. By the definition of PBE, the on-path beliefs satisfy $b(H \mid m_H, a) = \alpha\sigma^S(m_H \mid H)\rho / [\alpha\sigma^S(m_H \mid H)\rho + \beta\sigma^S(m_H \mid L)(1-\rho)]$ and $b(H \mid m_H, na) = (1-\alpha)\sigma^S(m_H \mid H)\rho / [(1-\alpha)\sigma^S(m_H \mid H)\rho + (1-\beta)\sigma^S(m_H \mid L)(1-\rho)]$. Given $\beta > \alpha$, $b(H \mid m_H, na) > b(H \mid m_H, a)$. The receiver's best responses are:

$$\sigma^R(r_H \mid m_H, a) \begin{cases} = 1 & \text{if } b(H \mid m_H, a)\Delta_H^R - (1 - b(H \mid m_H, a))\Delta_L^R > 0 \\ \in [0,1] & \text{if } b(H \mid m_H, a)\Delta_H^R - (1 - b(H \mid m_H, a))\Delta_L^R = 0 \\ = 0 & \text{if } b(H \mid m_H, a)\Delta_H^R - (1 - b(H \mid m_H, a))\Delta_L^R < 0 \end{cases}$$

$$\sigma^R(r_H \mid m_H, na) \begin{cases} = 1 & \text{if } b(H \mid m_H, na)\Delta_H^R - (1 - b(H \mid m_H, na))\Delta_L^R > 0 \\ \in [0,1] & \text{if } b(H \mid m_H, na)\Delta_H^R - (1 - b(H \mid m_H, na))\Delta_L^R = 0 \\ = 0 & \text{if } b(H \mid m_H, na)\Delta_H^R - (1 - b(H \mid m_H, na))\Delta_L^R < 0 \end{cases}$$

Because $b(H \mid m_H, na) > b(H \mid m_H, a)$, we have $b(H \mid m_H, a)\Delta_H^R - (1 - b(H \mid m_H, a))\Delta_L^R < b(H \mid m_H, na)\Delta_H^R - (1 - b(H \mid m_H, na))\Delta_L^R$. Therefore, $\sigma^R(r_H \mid m_H, a) \leq \sigma^R(r_H \mid m_H, na)$. □

We now prove that $\sigma^S(m_H \mid H) = 1$. Suppose, for contradiction, that there exists an equilibrium where $\sigma^S(m_L \mid H) > 0$. Then $\sigma^R(r_H \mid m_L, na)\Delta_H^S - C \geq \left(\alpha\sigma^R(r_H \mid m_H, a) + (1-\alpha)\sigma^R(r_H \mid m_H, na)\right)\Delta_H^S \Leftrightarrow -C \geq \left(\alpha\sigma^R(r_H \mid m_H, a) + (1-\alpha)\sigma^R(r_H \mid m_H, na) - \sigma^R(r_H \mid m_L, na)\right)\Delta_H^S$. This implies that:

$$\sigma^R(r_H \mid m_L, na) > \alpha\sigma^R(r_H \mid m_H, a) + (1-\alpha)\sigma^R(r_H \mid m_H, na) \tag{2}$$

We now show that a type L sender would get a lower expected payoff from sending m_H compared to m_L. If $\alpha = \beta$:

$$\begin{aligned} &\left(\beta\sigma^R(r_H \mid m_H, a) + (1-\beta)\sigma^R(r_H \mid m_H, na)\right)\Delta_L^S - C \\ =& \left(\alpha\sigma^R(r_H \mid m_H, a) + (1-\alpha)\sigma^R(r_H \mid m_H, na)\right)\Delta_L^S - C \\ \overset{(2)}{<}& \sigma^R(r_H \mid m_L, na)\Delta_L^S - C < \sigma^R(r_H \mid m_L, na)\Delta_L^S \end{aligned}$$

If $\alpha < \beta$:

$$\begin{aligned}
&\left(\beta\sigma^R(r_H \mid m_H, a) + (1-\beta)\sigma^R(r_H \mid m_H, na)\right)\Delta_L^S - C \\
\overset{Lemma\ 8}{\leq}\ &\left(\alpha\sigma^R(r_H \mid m_H, a) + (1-\alpha)\sigma^R(r_H \mid m_H, na)\right)\Delta_L^S - C \\
\overset{(2)}{<}\ &\sigma^R(r_H \mid m_L, na)\Delta_L^S - C < \sigma^R(r_H \mid m_L, na)\Delta_L^S
\end{aligned}$$

Therefore, a type L sender will always send m_L, meaning $\sigma^S(m_L \mid L) = 1$. Because $\sigma^S(m_L \mid L) = 1$, the receiver's expected payoff from taking r_L after observing $(m = m_L, l = na)$ is always higher than taking r_H: $[(\sigma^S(m_L \mid H)\rho)/(\sigma^S(m_L \mid H)\rho + 1 - \rho)]\Delta_H^R - [(1-\rho)/(\sigma^S(m_L \mid H)\rho + 1 - \rho)]\Delta_L^R \leq \rho\Delta_H^R - (1-\rho)\Delta_L^R < 0$. This means $\sigma^R(r_H \mid m_L, na) = 0$, which contradicts (2). Therefore, we must have $\sigma^S(m_H \mid H) = 1$. An immediate consequence is that if $m = m_L$ is an on-path message, the receiver will always take action r_L after receiving message m_L.

A.4 Proof of Lemma 2

When $\alpha = \beta = 0$, the lie detector is ineffective, and the receiver can only observe $\{m = m_H, l = na\}$ or $\{m = m_L, l = na\}$. The receiver's strategy is fully characterized by σ_{na}^R, the probability of taking r_H when there is no alarm. The best responses are:

$$\sigma_{na,\text{BR}}^R(\sigma^S) = \begin{cases} = 1, & \text{if } \sigma^S < \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \in [0,1], & \text{if } \sigma^S = \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ = 0, & \text{if } \sigma^S > \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \end{cases}; \sigma_{\text{BR}}^S(\sigma_{na}^R) = \begin{cases} = 1, & \text{if } \sigma_{na}^R > \frac{C}{\Delta_L^S} \\ \in [0,1], & \text{if } \sigma_{na}^R = \frac{C}{\Delta_L^S} \\ = 0, & \text{if } \sigma_{na}^R < \frac{C}{\Delta_L^S} \end{cases}$$

PBE requires that $\sigma^{S*} = \sigma_{\text{BR}}^S({\sigma_{na}^R}^*)$ and ${\sigma_{na}^R}^* = \sigma_{na,\text{BR}}^R(\sigma^{S*})$. This yields a unique equilibrium: $\sigma^{S*} = \rho\Delta_H^R/(1-\rho)\Delta_L^R$, ${\sigma_{na}^R}^* = C/\Delta_L^S$. In this equilibrium: $\mathbb{E}U^R = 0$, $\mathbb{E}U_L^S = 0$, and $\mathbb{E}U_H^S = C\Delta_H^S/\Delta_L^S$.

A.5 Proof of Lemma 4

When there are multiple equilibria, we select the Pareto-optimal equilibrium. The refinement does not drive the results because the area in the detector's capacity space $\{(0,\beta)|0 \leq \beta \leq 1\}$ with multiple equilibria, $\{(0,\beta)|\beta = \hat{\beta}\ or\ \beta = 1 - \rho\Delta_H^R/[(1-\rho)\Delta_L^R]\ and\ C < \rho\Delta_H^R/[(1-\rho)\Delta_L^R]\Delta_L^S\}$, has measure zero.

1. High lying cost $C \geq [\rho\Delta_H^R/[(1-\rho)\Delta_L^R]]\Delta_L^S$

$$\mathbb{E}U_L^S(\beta) = 0,\ \mathbb{E}U_H^S(\beta) = \begin{cases} \frac{C}{(1-\beta)\Delta_L^S}\Delta_H^S, & \beta < \hat{\beta} \\ \Delta_H^S, & \beta \geq \hat{\beta} \end{cases},\ \mathbb{E}U^R(\beta) = \begin{cases} 0, & \beta < \hat{\beta} \\ \rho\Delta_H^R, & \beta \geq \hat{\beta} \end{cases}$$

$$\mathbb{E}W(\beta) = \mathbb{E}U^R(\beta) + \rho\mathbb{E}U_H^S(\beta) + (1-\rho)\mathbb{E}U_L^S(\beta) = \begin{cases} \rho\frac{C}{(1-\beta)\Delta_L^S}\Delta_H^S, & \beta < \hat{\beta} \\ \rho(\Delta_H^S + \Delta_H^R), & \beta \geq \hat{\beta} \end{cases}$$

One can see that $\mathbb{E}U^R(\beta)$, $\mathbb{E}U_H^S(\beta)$, and $\mathbb{E}W(\beta)$ all (weakly) increase in β. They achieve the maximum value at any $\beta \geq \hat{\beta}$.

2. Low lying cost $C < [\rho\Delta_H^R/[(1-\rho)\Delta_L^R]]\Delta_L^S$

$$\mathbb{E}U_L^S(\beta) = \begin{cases} 0, & \beta < 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ (1-\beta)\Delta_L^S - C, & \beta \in \left[1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}, \hat{\beta}\right) \\ 0, & \beta \geq \hat{\beta} \end{cases}, \quad \mathbb{E}U_H^S(\beta) = \begin{cases} \frac{C}{(1-\beta)\Delta_L^S}\Delta_H^S, & \beta < 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \Delta_H^S, & \beta \geq 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \end{cases},$$

$$\mathbb{E}U^R(\beta) = \begin{cases} 0, & \beta < 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \rho\Delta_H^R - (1-\beta)(1-\rho)\Delta_L^R, & \beta \in \left[1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}, \hat{\beta}\right) \\ \rho\Delta_H^R, & \beta \geq \hat{\beta} \end{cases}$$

$$\mathbb{E}W(\beta) = \begin{cases} \rho\frac{C}{(1-\beta)\Delta_L^S}\Delta_H^S, & \beta < 1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ (1-\beta)(1-\rho)(\Delta_L^S - \Delta_L^R) + \rho(\Delta_H^S + \Delta_H^R) - (1-\rho)C, & \beta \in \left[1 - \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}, \hat{\beta}\right) \\ \rho(\Delta_H^S + \Delta_H^R), & \beta \geq \hat{\beta} \end{cases}$$

One can see that $\mathbb{E}U^R(\beta)$, $\mathbb{E}U_H^S(\beta)$, and $\mathbb{E}W(\beta)$ all (weakly) increase in β. $\mathbb{E}U^R(\beta)$ and $\mathbb{E}W(\beta)$ achieve the maximum value at any $\beta \geq \hat{\beta}$. $\mathbb{E}U_H^S(\beta)$ achieves the maximum value at any $\beta \geq 1 - \rho\Delta_H^R/[(1-\rho)\Delta_L^R]$.

A.6 Proof of Lemma 5

$$\alpha^*(\beta) = \min_{\lambda_L, \lambda_H \in [0,1]} \phi(s_L|\theta = H)\lambda_L + \phi(s_H|\theta = H)\lambda_H,$$
$$\text{s.t. } \phi(s_L|\theta = L)\lambda_L + \phi(s_H|\theta = L)\lambda_H = \beta$$

The constraint implies that

$$\lambda_L = \frac{\beta - \phi(s_H|\theta = L)\lambda_H}{\phi(s_L|\theta = L)} \tag{C1}$$

Substituting (C1) into the objective function, one can see that the coefficient of λ_H is positive:

$$\begin{aligned} &\phi(s_H|\theta = H) - \frac{\phi(s_H|\theta = L)\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)} \\ =&1 - \phi(s_L|\theta = H) - \frac{(1 - \phi(s_L|\theta = L))\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)} \\ =&1 - \frac{\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)} > 0 \end{aligned}$$

Therefore, the optimal λ_H^* with $\lambda_L = (\beta - \phi(s_H|\theta = L)\lambda_H)/\phi(s_L|\theta = L)$ should be the minimum feasible λ_H. The λ_H has restrictions: $(\beta - \phi(s_H|\theta = L)\lambda_H)/\phi(s_L|\theta = L) \in [0,1]$ and $\lambda_H \in [0,1]$. So, the minimum feasible λ_H is $\max\{(\beta - \phi(s_L|\theta = L))/\phi(s_H|\theta = L), 0\}$.

All in all,

$$\lambda_H^* = \max\left\{\frac{\beta - \phi(s_L|\theta = L)}{\phi(s_H|\theta = L)}, 0\right\}, \ \lambda_L^* = \frac{\beta}{\phi(s_L|\theta = L)} - \frac{\phi(s_H|\theta = L)}{\phi(s_L|\theta = L)}\lambda_1^*$$

If $\beta \leq \phi(s_L|\theta = L)$, then we have

$$\lambda_H^* = 0, \lambda_L^* = \frac{\beta}{\phi(s_L|\theta = L)}, \alpha^*(\beta) = \frac{\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)}\beta$$

If $\beta > \phi(s_L|\theta = L)$, then we have

$$\lambda_H^* = \frac{\beta - \phi(s_L|\theta = L)}{\phi(s_H|\theta = L)}, \lambda_L^* = 1$$

$$\begin{aligned}\alpha^*(\beta) &= \phi(s_L|\theta = H) + \phi(s_H|\theta = H)\frac{\beta - \phi(s_L|\theta = L)}{\phi(s_H|\theta = L)} \\ &= \frac{\phi(s_H|\theta = H)}{\phi(s_H|\theta = L)}\beta + \left(1 - \frac{\phi(s_H|\theta = H)}{\phi(s_H|\theta = L)}\right)\end{aligned}$$

A.7 Proof of Lemma 6

We start with two general lemmas to establish the optimization principle, then apply it to $J(\beta, \alpha)$.

Lemma 9. *Let $f_1, f_2, f_3 : \mathcal{F} \to \mathbb{R}$ be real-valued functions on the feasible set $\mathcal{F} \subset \mathcal{A} \times \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are non-empty sets, and $\mathcal{F}$ is defined by constraints coupling $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$. For each $\beta \in \mathcal{B}$, let $A(\beta) = \{\alpha \mid (\beta, \alpha) \in \mathcal{F}\}$ be non-empty, and assume there exists a common $\alpha^*(\beta) \in A(\beta)$ such that:*

$$f_i(\beta, \alpha^*(\beta)) = \max_{\alpha \in A(\beta)} f_i(\beta, \alpha) \quad \textit{for } i = 1, 2, 3,$$

with maxima attained. Define the weighted function $f(\beta, \alpha) = w_1 f_1(\beta, \alpha) + w_2 f_2(\beta, \alpha) + w_3 f_3(\beta, \alpha)$, where $w_1, w_2, w_3 \geq 0$. Then:

$$\max_{(\beta,\alpha) \in \mathcal{F}} f(\beta, \alpha) = \max_{\beta \in \mathcal{B}} f(\beta, \alpha^*(\beta)).$$

Proof. Define $g : \mathcal{B} \to \mathbb{R}$ by $g(\beta) = f(\beta, \alpha^*(\beta)) = w_1 f_1(\beta, \alpha^*(\beta)) + w_2 f_2(\beta, \alpha^*(\beta)) + w_3 f_3(\beta, \alpha^*(\beta))$, where $(\beta, \alpha^*(\beta)) \in \mathcal{F}$. We prove $\max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha) = \max_{\beta \in \mathcal{B}} g(\beta)$.

- $\max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha) \geq \max_{\beta \in \mathcal{B}} g(\beta)$. For any $\beta \in \mathcal{B}$, since $(\beta, \alpha^*(\beta)) \in \mathcal{F}$, we have $f(\beta, \alpha^*(\beta)) = g(\beta) \leq \max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha)$. Thus, $\max_{\beta \in \mathcal{B}} g(\beta) \leq \max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha)$.
- $\max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha) \leq \max_{\beta \in \mathcal{B}} g(\beta)$. For any $(\beta, \alpha) \in \mathcal{F}$, since $\alpha \in A(\beta)$ and $\alpha^*(\beta)$ maximizes each $f_i(\cdot, \beta)$ over $A(\beta)$, we have $f_i(\beta, \alpha) \leq f_i(\beta, \alpha^*(\beta))$ for $i = 1, 2, 3$. Since $w_i \geq 0$, it follows that:
$$f(\beta, \alpha) = \sum_{i=1}^{3} w_i f_i(\beta, \alpha) \leq \sum_{i=1}^{3} w_i f_i(\beta, \alpha^*(\beta)) = f(\beta, \alpha^*(\beta)) = g(\beta).$$
Since $g(\beta) \leq \max_{\beta' \in \mathcal{B}} g(\beta')$, we have $f(\beta, \alpha) \leq \max_{\beta' \in \mathcal{B}} g(\beta')$. This holds for all $(\beta, \alpha) \in \mathcal{F}$, so $\max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha) \leq \max_{\beta \in \mathcal{B}} g(\beta)$.

Thus, $\max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha) = \max_{\beta \in \mathcal{B}} g(\beta)$. □

Lemma 10. *Let $f_1, f_2, f_3 : \mathcal{F} \to \mathbb{R}$ be real-valued functions on the feasible set $\mathcal{F} \subset \mathcal{A} \times \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are non-empty sets, and $\mathcal{F}$ is defined by constraints coupling $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$. For each $\beta \in \mathcal{B}$, let $A(\beta) = \{\alpha \mid (\beta, \alpha) \in \mathcal{F}\}$ be non-empty, and assume there exists a common $\alpha^*(\beta) \in A(\beta)$ such that:*

$$f_i(\beta, \alpha^*(\beta)) = \max_{\alpha \in A(\beta)} f_i(\beta, \alpha) \quad \textit{for } i = 1, 2, 3,$$

with maxima attained. Define the weighted function $f(\beta, \alpha) = w_1 f_1(\beta, \alpha) + w_2 f_2(\beta, \alpha) + w_3 f_3(\beta, \alpha)$, where $w_1, w_2, w_3 \geq 0$, and assume that for each $\beta \in \mathcal{B}$, there exists a unique $\alpha \in A(\beta)$ that maximizes $f(\beta, \alpha)$. Then:

$$\left\{(\beta, \alpha^*(\beta)) \mid \beta \in \mathcal{B},\ f(\beta, \alpha^*(\beta)) = \max_{\beta' \in \mathcal{B}} f(\alpha^*(\beta'), \beta')\right\} = \left\{(\beta, \alpha) \mid (\beta, \alpha) \in \mathcal{F},\ f(\beta, \alpha) = \max_{(\beta', \alpha') \in \mathcal{F}} f(\beta', \alpha')\right\}.$$

Proof. Define $M = \max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha)$ and $g(\beta) = f(\beta, \alpha^*(\beta))$, where $(\beta, \alpha^*(\beta)) \in \mathcal{F}$. By Lemma 9, $M = \max_{\beta \in \mathcal{B}} g(\beta)$. Let:

- $S_1 = \{(\beta, \alpha^*(\beta)) \mid \beta \in \mathcal{B},\ g(\beta) = \max_{\beta' \in \mathcal{B}} g(\beta')\}$,
- $S_2 = \{(\beta, \alpha) \mid (\beta, \alpha) \in \mathcal{F},\ f(\beta, \alpha) = \max_{(\beta', \alpha') \in \mathcal{F}} f(\beta', \alpha')\}$.

We prove $S_1 = S_2$.

Step 1: $S_1 \subseteq S_2$**.** For $(\beta, \alpha^*(\beta)) \in S_1$, we have $g(\beta) = f(\beta, \alpha^*(\beta)) = \max_{\beta' \in \mathcal{B}} g(\beta') = M$. Since $(\beta, \alpha^*(\beta)) \in \mathcal{F}$ and $f(\beta, \alpha^*(\beta)) = M$, it follows that $(\beta, \alpha^*(\beta)) \in S_2$.

Step 2: $S_2 \subseteq S_1$**.** For $(\beta, \alpha) \in S_2$, we have $(\beta, \alpha) \in \mathcal{F}$ and $f(\beta, \alpha) = M$. Since $\alpha^*(\beta)$ maximizes each $f_i(\cdot, \beta)$ over $A(\beta)$, and $w_i \geq 0$, we have:

$$f(\beta, \alpha) = \sum_{i=1}^{3} w_i f_i(\beta, \alpha) \leq \sum_{i=1}^{3} w_i f_i(\beta, \alpha^*(\beta)) = f(\beta, \alpha^*(\beta)).$$

Since $f(\beta, \alpha) = M$ and $f(\beta, \alpha^*(\beta)) \leq M$, we get $f(\beta, \alpha^*(\beta)) = M$. By the uniqueness of the maximizer of $f(\cdot, \beta)$ over $A(\beta)$, $\alpha = \alpha^*(\beta)$. Thus, $f(\beta, \alpha^*(\beta)) = M = \max_{\beta' \in \mathcal{B}} g(\beta')$, so $(\beta, \alpha) = (\beta, \alpha^*(\beta)) \in S_1$.

Since $S_1 \subseteq S_2$ and $S_2 \subseteq S_1$, we conclude $S_1 = S_2$. □

Application to the Optimal Design Problem

Consider the following optimal design problem:

$$\begin{aligned} \max_{(\beta,\alpha)} \quad & J(\beta, \alpha) = w_R \mathbb{E}U^R(\beta, \alpha) + w_H \mathbb{E}U_H^S(\beta, \alpha) + w_L \mathbb{E}U_L^S(\beta, \alpha) \\ \text{s.t.} \quad & (\beta, \alpha) \in \mathcal{F}(\phi), \end{aligned}$$

where the feasible set $\mathcal{F}(\phi)$ is defined as:

$$\mathcal{F}(\phi) = \left\{ (\beta, \alpha) \middle| \begin{array}{l} \beta = \phi(s_L|\theta = L)\lambda_L + \phi(s_H|\theta = L)\lambda_H, \\ \alpha = \phi(s_L|\theta = H)\lambda_L + \phi(s_H|\theta = H)\lambda_H, \\ \lambda_L, \lambda_H \in [0,1] \end{array} \right\}.$$

By Proposition 2, for each $\beta \in \mathcal{B}$, there exists $\alpha^*(\beta;\phi) \in A(\beta)$ such that:

$$\begin{aligned} \mathbb{E}U^R(\beta, \alpha^*(\beta;\phi)) &= \max_{\alpha \in A(\beta)} \mathbb{E}U^R(\beta, \alpha), \\ \mathbb{E}U_H^S(\beta, \alpha^*(\beta;\phi)) &= \max_{\alpha \in A(\beta)} \mathbb{E}U_H^S(\beta, \alpha), \\ \mathbb{E}U_L^S(\beta, \alpha^*(\beta;\phi)) &= \max_{\alpha \in A(\beta)} \mathbb{E}U_L^S(\beta, \alpha). \end{aligned}$$

We select the Pareto-optimal equilibrium for those cases with multiple equilibria, thus for each $\beta \in \mathcal{B}$, the maximizer $\alpha^*(\beta) \in A(\beta)$ of $J(\beta, \alpha)$ is unique.

Step 1: Maximum Value. Applying Lemma 9 with $f_1 = \mathbb{E}U^R$, $f_2 = \mathbb{E}U_H^S$, $f_3 = \mathbb{E}U_L^S$, and weights $w_R, w_H, w_L \geq 0$, we have $\max_{(\beta,\alpha)\in\mathcal{F}(\phi)} J(\beta,\alpha) = \max_{\beta\in\mathcal{B}} J(\beta, \alpha^*(\beta;\phi))$.

Step 2: Maximizers. Applying Lemma 10, we obtain the set of maximizers as:

$$\begin{aligned} & \left\{ (\beta, \alpha^*(\beta;\phi)) \middle| \beta \in \mathcal{B},\ J(\beta, \alpha^*(\beta;\phi)) = \max_{\beta' \in \mathcal{B}} J(\beta', \alpha^*(\beta')) \right\} \\ = & \left\{ (\beta, \alpha) \middle| (\beta, \alpha) \in \mathcal{F}(\phi),\ J(\beta, \alpha) = \max_{(\beta', \alpha') \in \mathcal{F}(\phi)} J(\beta', \alpha') \right\}. \end{aligned}$$

A.8 Proof of Proposition 3

For simplicity, we slightly abuse the notation by denoting $U^R(\beta, \alpha^*(\beta;\phi))$ by $U^R(\beta)$, $U_H^S(\beta, \alpha^*(\beta;\phi))$ by $U_H^S(\beta)$, and $U_L^S(\beta, \alpha^*(\beta;\phi))$ by $U_L^S(\beta)$.

We first characterize the sender's equilibrium strategy for a given detector $\{\beta, \alpha^*(\beta;\phi)\}$.

Lemma 11. *For a classifier with a high capacity, the equilibrium is*

$$\begin{cases} \sigma^S = \frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in [0, \beta_1) \\ \sigma^S = 1, \sigma_{na}^R = 1, \sigma_a^R = 0, & \beta \in \left[\beta_1, \hat{\beta}\right), \\ \sigma^S = \frac{\alpha^*(\beta;\phi)\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

where $\beta_1 := [\rho\Delta_H^R - (1-\rho)\Delta_L^R]/[(\phi(s_L|\theta = H)/\phi(s_L|\theta = L))\rho\Delta_H^R - (1-\rho)\Delta_L^R] \leq \hat{\beta}$.

For a classifier with a low capacity, the equilibrium is

$$\begin{cases} \sigma^S = \frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in \left[0, \hat{\beta}\right) \\ \sigma^S = \frac{\alpha^*(\beta;\phi)\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

Proof. The equilibrium when $\beta \geq \hat{\beta}$ follows directly from Proposition 1. Now consider $\beta \in [0, \hat{\beta})$, Proposition 1 implies that the equilibrium is $\{\sigma^S = 1, \sigma^R_{na} = 1, \sigma^R_a = 0\}$ if $(1-\alpha^*(\beta;\phi))/(1-\beta) \geq ((1-\rho)\Delta^R_L)/(\rho\Delta^R_H)$ and is $\{\sigma^S = [(1-\alpha^*(\beta;\phi))\rho\Delta^R_H]/[(1-\beta)(1-\rho)\Delta^R_L], \sigma^R_{na} = C/((1-\beta)\Delta^S_L), \sigma^R_a = 0\}$ if $(1-\alpha^*(\beta;\phi))/(1-\beta) < [(1-\rho)\Delta^R_L]/(\rho\Delta^R_H)$.

According to Lemma 5,

$$\begin{aligned}\frac{1-\alpha^*(\beta;\phi)}{1-\beta} &= \begin{cases} -\frac{\beta}{1-\beta}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} + \frac{1}{1-\beta}, & if\ \beta \leq \phi(s_L|\theta=L) \\ \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} & if\ \beta > \phi(s_L|\theta=L), \end{cases} \\ &= \min\left\{-\frac{\beta}{1-\beta}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} + \frac{1}{1-\beta}, \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\right\} \in \left[1, \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\right]\end{aligned}$$

Let $g(\beta) := -[\beta/(1-\beta)][\phi(s_L|\theta=H)/\phi(s_L|\theta=L)] + 1/(1-\beta)$, which increases in β. One can see that $g(\hat{\beta}) = -[(\Delta^S_L - C)/C][\phi(s_L|\theta=H)/\phi(s_L|\theta=L)] + \Delta^S_L/C$.

- If the classifier has a low capacity, $\phi(s_H|\theta=H)/\phi(s_H|\theta=L) < (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$ or $\phi(s_L|\theta=L)/\phi(s_L|\theta=H) < (\Delta^S_L - C)\rho\Delta^R_H/[\Delta^S_L\rho\Delta^R_H - (1-\rho)\Delta^R_L C]$(*i.e.* $g(\hat{\beta}) < (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$), then $(1-\alpha^*(\beta;\phi))/(1-\beta) < (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$ and the equilibrium is $\{\sigma^S = [(1-\alpha^*(\beta;\phi))\rho\Delta^R_H]/[(1-\beta)(1-\rho)\Delta^R_L], \sigma^R_{na} = C/[(1-\beta)\Delta^S_L], \sigma^R_a = 0\}$ for all $\beta < \hat{\beta}$.

- If the classifier has a high capacity, $\phi(s_H|\theta=H)/\phi(s_H|\theta=L) \geq (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$ and $\phi(s_L|\theta=L)/\phi(s_L|\theta=H) \geq (\Delta^S_L - C)\rho\Delta^R_H/[\Delta^S_L\rho\Delta^R_H - (1-\rho)\Delta^R_L C]$(*i.e.* $g(\hat{\beta}) \geq (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$), then $\beta_1 \in \left(0, \hat{\beta}\right]$, $g(\beta_1) = (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$, and the equilibrium is

$$\begin{cases} \sigma^S = \frac{(1-\alpha^*(\beta;\phi))\rho\Delta^R_H}{(1-\beta)(1-\rho)\Delta^R_L}, \sigma^R_{na} = \frac{C}{(1-\beta)\Delta^S_L}, \sigma^R_a = 0, & \beta \in [0, \beta_1) \\ \sigma^S = 1, \sigma^R_{na} = 1, \sigma^R_a = 0, & \beta \in \left[\beta_1, \hat{\beta}\right) \end{cases}$$

□

According to Lemma 11 and Table 3, $\mathbb{E}U^R(\beta)$ weakly decreases in β when $\beta \in \left[\hat{\beta}, 1\right)$.

1. If the classifier has a low capacity, $\mathbb{E}U^R(\beta) = 0$ when $\beta < \hat{\beta}$. When $\beta \geq \hat{\beta}$, $\mathbb{E}U^R(\beta) = [1 - \alpha^*(\beta;\phi)/\beta]\rho\Delta^R_H$, which is constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta=L)\}]$ and strictly decreases in β for $\beta > \max\{\hat{\beta}, \phi(s_L|\theta=L)\}$. So, the optimal true-positive rate of the detector that maximizes the receiver's expected payoff is any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta=L)\}]$.

2. If the classifier has a high capacity, then

$$\mathbb{E}U^R(\beta) = \begin{cases} 0, & \beta \in [0, \beta_1) \\ (1-\alpha^*(\beta;\phi))\rho\Delta^R_H - (1-\beta)(1-\rho)\Delta^R_L, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \left(1 - \frac{\alpha^*(\beta;\phi)}{\beta}\right)\rho\Delta^R_H, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

Lemma 12. *$\phi(s_L|\theta=L) \geq \beta_1$ if and only if $\phi(s_H|\theta=H)/\phi(s_H|\theta=L) \geq (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$.*

Proof. Because $\mathcal{Z}(x) := (x-1)/[x - \phi(s_L|\theta = H)/\phi(s_L|\theta = L)]$ increases in x, $\phi(s_L|\theta = L) := \mathcal{Z}\left(\phi(s_H|\theta = H)/\phi(s_H|\theta = L)\right) \geq \beta_1 := \mathcal{Z}\left((1-\rho)\Delta_L^R/(\rho\Delta_H^R)\right)$ if and only if $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) \geq (1-\rho)\Delta_L^R/(\rho\Delta_H^R)$. □

(a) If $\hat{\beta} > \phi(s_L|\theta = L)$,

$$\frac{\partial \mathbb{E}U^R(\beta)}{\partial \beta} = \begin{cases} (1-\rho)\Delta_L^R - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\rho\Delta_H^R > 0, & \beta \in [\beta_1, \phi(s_L|\theta = L)) \\ (1-\rho)\Delta_L^R - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\rho\Delta_H^R \leq 0, & \beta \in \left[\phi(s_L|\theta = L), \hat{\beta}\right) \end{cases}$$

So, $\mathbb{E}U^R(\beta)$ is maximized at $\phi(s_L|\theta = L)$ for $\beta \in [\beta_1, \hat{\beta})$. Because $\mathbb{E}U^R(\beta) = 0$ for all $\beta \leq \beta_1$ and $\mathbb{E}U^R(\beta)$ is maximized at $\hat{\beta}$ for $\beta \in [\hat{\beta}, 1)$, the maximizer of $\mathbb{E}U^R(\beta)$ among $\beta \in [0,1]$ must be $\hat{\beta}$ or $\phi(s_L|\theta = L)$.[1] Thus, we only need to compare $\mathbb{E}U^R(\hat{\beta})$ and $\mathbb{E}U^R(\phi(s_L|\theta = L))$. One can see that $\mathbb{E}U^R(\hat{\beta}) = (1/\hat{\beta} - 1)\left(\phi(s_H|\theta = H)/\phi(s_H|\theta = L) - 1\right)\rho\Delta_H^R$ and $\mathbb{E}U^R(\phi(s_L|\theta = L)) = \phi(s_H|\theta = H)\rho\Delta_H^R - \phi(s_H|\theta = L)(1-\rho)\Delta_L^R = \phi(s_H|\theta = L)[\rho\Delta_H^R\phi(s_H|\theta = H)/\phi(s_H|\theta = L) - (1-\rho)\Delta_L^R]$. So, $\mathbb{E}U^R(\hat{\beta}) \geq \mathbb{E}U^R(\phi(s_L|\theta = L))$ if and only if:

$$\hat{\beta} \leq \frac{\phi(s_H \mid \theta = H) - \phi(s_H \mid \theta = L)}{\left[\phi(s_H|\theta = H) - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H \mid \theta = L)\right]\phi(s_H \mid \theta = L) + \phi(s_H \mid \theta = H) - \phi(s_H \mid \theta = L)}$$
$$=: \hat{\beta}_{\text{critical}}, \text{ which increases in } \phi(s_L|\theta = H)/\phi(s_L|\theta = L).$$

(b) If $\hat{\beta} \leq \phi(s_L|\theta = L)$,

$$\frac{\partial \mathbb{E}U^R(\beta)}{\partial \beta} = (1-\rho)\Delta_L^R - \frac{\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)}\rho\Delta_H^R > 0, \beta \in \left[\beta_1, \hat{\beta}\right)$$

Hence, all $\beta \in \left[\beta_1, \hat{\beta}\right)$ is dominated by $\beta = \hat{\beta}$. We can also find $\mathbb{E}U^R(\beta)$ is constant for $\beta \in [\hat{\beta}, \phi(s_L|\theta = L)]$ and decreasing for $\beta > \phi(s_L|\theta = L)$.

All in all, if the classifier has a high capacity, the optimal true-positive rate for the receiver is any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ if $\hat{\beta} \leq \hat{\beta}_{\text{critical}}$ and is $\beta = \phi(s_L|\theta = L)$ if $\hat{\beta} > \hat{\beta}_{\text{critical}}$.

$$\hat{\beta} \leq \hat{\beta}_{\text{critical}}$$
$$\Leftrightarrow\ C \geq \hat{C} = \frac{\left[\phi(s_H|\theta = H) - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H \mid \theta = L)\right]\phi(s_H \mid \theta = L)}{\left[\phi(s_H|\theta = H) - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H \mid \theta = L) - 1\right]\phi(s_H \mid \theta = L) + \phi(s_H \mid \theta = H)}\Delta_L^S.$$

Therefore, if the classifier has a high capacity, the optimal true-positive rate for the receiver is any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ if $C \geq \hat{C}$ and is $\beta = \phi(s_L|\theta = L)$ if $C < \hat{C}$.

[1] If $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) = (1-\rho)\Delta_L^R/(\rho\Delta_H^R)$, $\phi(s_L|\theta = L) = \beta_1$ and choosing $\beta = \phi(s_L|\theta = L)$ is equivalent to choosing any $\beta < \beta_1$ which makes zero payoff and is dominated by choosing $\beta = \hat{\beta}$.

A.9 Proof of Proposition 4

1. Classifier with a low capacity

   According to Lemma 11, the equilibria are

$$\begin{cases} \sigma^S = \frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in \left[0, \hat{\beta}\right) \\ \sigma^S = \frac{\alpha^*(\beta;\phi)\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

   So, the payoff of the low-type sender is $\mathbb{E}U_L^S(\beta) = 0$ and the payoff of the high-type sender is

$$\mathbb{E}U_H^S(\beta) = \begin{cases} C\frac{\Delta_H^S}{\Delta_L^S}\frac{1-\alpha^*(\beta;\phi)}{1-\beta}, & \beta \in \left[0, \hat{\beta}\right) \\ \Delta_H^S - (\Delta_L^S - C)\frac{\Delta_H^S}{\Delta_L^S}\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

   For $\beta \in [0, \hat{\beta})$, $\mathbb{E}U_H^S(\beta)$ is weakly increasing and is dominated by $\beta = \hat{\beta}$. $\mathbb{E}U_H^S(\beta)$ is constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ and is decreasing for $\beta > \max\{\hat{\beta}, \phi(s_L|\theta = L)\}$. So, in this case, $\mathbb{E}U_H^S(\beta)$ is maximized at $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$.

2. Classifier with a high capacity

   According to Lemma 11, the equilibria are

$$\begin{cases} \sigma^S = \frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in [0, \beta_1) \\ \sigma^S = 1, \sigma_{na}^R = 1, \sigma_a^R = 0, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \sigma^S = \frac{\alpha^*(\beta;\phi)\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

   So, the payoff of the low-type sender is

$$\mathbb{E}U_L^S(\beta) = \begin{cases} 0, & \beta \in [0, \beta_1) \\ (1-\beta)\Delta_L^S - C & \beta \in \left[\beta_1, \hat{\beta}\right) \\ 0, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

   and the utility of the high-type sender is

$$\mathbb{E}U_H^S(\beta) = \begin{cases} C\frac{\Delta_H^S}{\Delta_L^S}\frac{1-\alpha^*(\beta;\phi)}{1-\beta} & \beta \in [0, \beta_1) \\ (1-\alpha^*(\beta;\phi))\Delta_H^S, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \Delta_H^S - (\Delta_L^S - C)\frac{\Delta_H^S}{\Delta_L^S}\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

   Note that we have proved $\phi(s_L|\theta = L) \geq \beta_1$ for a strong classifier in Lemma 12. Hence, for $\beta \in [0, \beta_1)$, $\mathbb{E}U_H^S(\beta)$ is increasing in β. Moreover, $\mathbb{E}U_H^S(\beta)$ is decreasing in $\beta \in \left[\beta_1, \hat{\beta}\right)$, constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$, and decreasing for $\beta > \max\{\hat{\beta}, \phi(s_L|\theta = L)\}$.

Because $\mathbb{E}U_H^S(\hat{\beta}) = \Delta_H^S - (\Delta_L^S - C)(\Delta_H^S/\Delta_L^S)(\alpha^*(\hat{\beta};\phi)/\hat{\beta}) = (1-\alpha^*(\hat{\beta};\phi))\Delta_H^S \leq (1-\alpha^*(\beta_1;\phi))\Delta_H^S = \mathbb{E}U_H^S(\beta_1)$, $\mathbb{E}U_H^S(\beta)$ is maximized at β_1. One can see that $\mathbb{E}U_L^S(\beta)$ is also maximized at β_1.

To show that $\beta_1 = [\rho\Delta_H^R - (1-\rho)\Delta_L^R]/[(\phi(s_L|\theta=H)/\phi(s_L|\theta=L))\rho\Delta_H^R - (1-\rho)\Delta_L^R]$ decreases in $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)$, one just need to observe that the numerator of β_1 is negative, $(1-\rho)\Delta_L^R > 0$, and $\rho\Delta_H^R > 0$.

A.10 Proof of Proposition 5

We begin by establishing that $\mathbb{E}\widetilde{W}(\beta)$ is equivalent to $\mathbb{E}W(\beta)$ with the following parameter rescaling:

$$\Delta_H'^S = w_H\Delta_H^S, \quad \Delta_L'^S = w_L\Delta_L^S, \quad C' = w_LC, \quad \hat{C}' = w_L\hat{C}.$$

The expected social welfare function can be expressed as:

$$\begin{aligned}\mathbb{E}\widetilde{W}(\beta) &= \mathbb{E}U^R(\beta) + w_L(1-\rho)\mathbb{E}U_L^S(\beta) + \rho w_H\mathbb{E}U_H^S(\beta)\\ &= \begin{cases} \rho C'\frac{\Delta_H'^S}{\Delta_L'^S}\frac{1-\alpha^*(\beta;\phi)}{1-\beta} & \beta\in[0,\beta_1)\\ \rho(1-\alpha^*(\beta;\phi))(\Delta_H'^S+\Delta_H^R) + (1-\rho)\left[(1-\beta)(\Delta_L'^S-\Delta_L^R) - C'\right] & \beta\in\left[\beta_1,\hat{\beta}\right)\\ \rho\left(1-\frac{\alpha^*(\beta;\phi)}{\beta}\right)(\Delta_H'^S+\Delta_H^R) + \rho C'\frac{\Delta_H'^S}{\Delta_L'^S}\frac{\alpha^*(\beta;\phi)}{\beta} & \beta\in\left[\hat{\beta},1\right)\end{cases}\end{aligned}$$

The proof proceeds by analyzing two distinct cases based on the classifier's capacity.

Case 1: Low-capacity classifier. According to Propositions 3 and 4, $\beta\in[\hat{\beta},\max\{\hat{\beta},\phi(s_L|\theta=L)\}]$ constitutes the optimal detector for both the receiver and the sender. Consequently, any β in this interval maximizes social welfare when the classifier has low capacity.

Case 2: High-capacity classifier. From Propositions 3 and 4, we obtain the following optimal lie detector configurations:

- *Receiver's optimization:* $\{\beta^*\in[\hat{\beta},\max\{\hat{\beta},\phi(s_L|\theta=L)\}], \alpha^*=\alpha^*(\beta^*;\phi)\}$ when $\hat{\beta}\leq\hat{\beta}_{\text{critical}}$, and $\{\beta^*=\phi(s_L\mid\theta=L), \alpha^*=\alpha^*(\beta^*;\phi)\}$ when $\hat{\beta}>\hat{\beta}_{\text{critical}}$.
- *Sender's optimization:* $\{\beta^*=\beta_1, \alpha^*=\alpha^*(\beta^*;\phi)\}$.

This yields the following expression for expected social welfare:

$$\mathbb{E}\widetilde{W}(\beta) = \begin{cases} \underbrace{\mathbb{E}U^R(\beta)}_{\substack{\text{maximized at any}\\ \beta\in[\hat{\beta},\max\{\hat{\beta},\phi(s_L|\theta=L)\}]}} + \underbrace{w_L(1-\rho)\mathbb{E}U_L^S(\beta) + w_H\rho\mathbb{E}U_H^S(\beta)}_{\text{maximized at }\beta=\beta_1}, & \text{if } \hat{\beta}\leq\hat{\beta}_{\text{critical}}\\ \underbrace{\mathbb{E}U^R(\beta)}_{\text{maximized at }\beta=\phi(s_L|\theta=L)} + \underbrace{w_L(1-\rho)\mathbb{E}U_L^S(\beta) + w_H\rho\mathbb{E}U_H^S(\beta)}_{\text{maximized at }\beta=\beta_1}, & \text{if } \hat{\beta}>\hat{\beta}_{\text{critical}}\end{cases}$$

First, a high-capacity classifier satisfies $\phi(s_L|\theta=L)/\phi(s_L|\theta=H) \geq (\Delta_L^S - C)\rho\Delta_H^R/[\Delta_L^S\rho\Delta_H^R - (1-\rho)\Delta_L^RC]$, which is equivalent to $C\leq\Delta_L^S(1-\beta_1)$.

To characterize the optimal true-positive rate, we analyze the following three scenarios based on the lying cost parameter:

1. **Low lying cost regime** ($C < \hat{C}$): This condition is equivalent to $\hat{\beta} > \hat{\beta}_{\text{critical}}$. In this case, the optimal β lies in the interval $[\beta_1, \phi(s_L \mid \theta = L)]$.
2. **Intermediate lying cost regime** ($C \in [\hat{C}, \Delta_L^S \phi(s_H \mid \theta = L))$): This corresponds to $\hat{\beta} \in (\phi(s_L \mid \theta = L), \hat{\beta}_{\text{critical}}]$. The optimal β lies in the interval $[\beta_1, \hat{\beta}]$.
3. **High lying cost regime** ($C \in [\Delta_L^S \phi(s_H \mid \theta = L), \Delta_L^S(1 - \beta_1)]$): This corresponds to $\hat{\beta} \leq \phi(s_L \mid \theta = L)$. The optimal β lies in the interval $[\beta_1, \phi(s_L \mid \theta = L)]$.

The inequality $\hat{C} < \Delta_L^S \phi(s_H \mid \theta = L)$ implies $\hat{\beta}_{\text{critical}} > \phi(s_L \mid \theta = L)$.

A.10.1 Low lying cost regime: $C < \hat{C}$

In this regime, we have $\hat{\beta} > \hat{\beta}_{\text{critical}} > \phi(s_L \mid \theta = L)$, which implies that we need only consider $\beta \in [\beta_1, \phi(s_L \mid \theta = L)]$. For this interval, the expected social welfare function takes the following form:

$$\mathbb{E}\widetilde{W}(\beta) = \rho \left(1 - \frac{\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)}\beta\right)(\Delta_H'^S + \Delta_H^R) + (1 - \rho)\left[(1 - \beta)(\Delta_L'^S - \Delta_L^R) - C'\right]$$

Because $\mathbb{E}\widetilde{W}(\beta)$ is linear in β over the interval $[\beta_1, \phi(s_L \mid \theta = L)]$, the optimal true-positive rate β^* is characterized by:

$$\beta^* = \begin{cases} \phi(s_L \mid \theta = L) & \text{if } \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}(\Delta_H'^S + \Delta_H^R) + (1-\rho)(\Delta_L'^S - \Delta_L^R) < 0 \\ \beta_1 & \text{if } \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}(\Delta_H'^S + \Delta_H^R) + (1-\rho)(\Delta_L'^S - \Delta_L^R) > 0 \\ \text{any value in } [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}(\Delta_H'^S + \Delta_H^R) + (1-\rho)(\Delta_L'^S - \Delta_L^R) = 0 \end{cases}$$

To simplify the characterization of the optimal β, we define two parameters:

$$n_0 := \frac{\rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^S}{(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R} > 0, \;\; l_0 \;\; := \frac{(1-\rho)\Delta_L^S}{(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R} > 0.$$

Using these parameters, we can express the set of optimal true-positive rates more concisely as:

$$\mathcal{B}(w_H, w_L) := \begin{cases} \{\phi(s_L \mid \theta = L)\} & \text{if } n_0 w_H + l_0 w_L < 1 \\ \{\beta_1\} & \text{if } n_0 w_H + l_0 w_L > 1 \\ [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } n_0 w_H + l_0 w_L = 1 \end{cases}$$

and the maximum expected welfare can be written as:

$$\max_{\beta\in[\beta_1,\phi(s_L|\theta=L)]} \mathbb{E}\widetilde{W}(\beta) = \mathbb{E}\widetilde{W}(\phi(s_L \mid \theta = L)) + (n_0 w_H + l_0 w_L - 1)^+ \left[(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^R\right](\phi(s_L \mid \theta = L) - \beta_1)$$

A.10.2 Intermediate lying cost regime: $C \in [\hat{C}, \Delta_L^S \phi(s_H \mid \theta = L))$

In this regime, we have $\hat{\beta} \in (\phi(s_L \mid \theta = L), \hat{\beta}_{\text{critical}}]$, which implies that we need only consider $\beta \in [\beta_1, \hat{\beta}]$. For this interval, the expected social welfare function takes the following form for $\beta \in [\beta_1, \hat{\beta}]$:

$$\mathbb{E}\widetilde{W}(\beta) = \begin{cases} \rho(1 - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\beta)(\Delta_H'^S + \Delta_H^R) + (1-\rho)\left[(1-\beta)(\Delta_L'^S - \Delta_L^R) - C'\right] & \beta \in [\beta_1, \phi(s_L \mid \theta = L)) \\ \rho(1-\beta)\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}(\Delta_H'^S + \Delta_H^R) + (1-\rho)\left[(1-\beta)(\Delta_L'^S - \Delta_L^R) - C'\right] & \beta \in \left[\phi(s_L \mid \theta = L), \hat{\beta}\right) \\ -\rho\frac{C}{\Delta_L^S - C}\left(1 - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\right)\Delta_H^R + \rho C\frac{\Delta_H'^S}{\Delta_L^S}\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}, & \beta = \hat{\beta} \end{cases}$$

1. For $\beta \in [\beta_1, \phi(s_L \mid \theta = L)]$, the pattern of $\mathbb{E}\widetilde{W}(\beta)$ is the same as the low lying cost regime. That is,

$$\operatorname*{argmax}_{\beta\in[\beta_1,\phi(s_L|\theta=L)]} \mathbb{E}\widetilde{W}(\beta) = \mathcal{B}(w_H, w_L) := \begin{cases} \{\phi(s_L \mid \theta = L)\} & \text{if } n_0 w_H + l_0 w_L < 1 \\ \{\beta_1\} & \text{if } n_0 w_H + l_0 w_L > 1 \\ [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } n_0 w_H + l_0 w_L = 1 \end{cases}$$

2. For $\beta \in \left[\phi(s_L \mid \theta = L), \hat{\beta}\right)$, we have

$$\begin{aligned} \frac{\partial \mathbb{E}\widetilde{W}(\beta)}{\partial \beta} &= -\rho\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}(\Delta_H'^S + \Delta_H^R) - (1-\rho)(\Delta_L'^S - \Delta_L^R) \\ &= \underbrace{(1-\rho)\Delta_L^R - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\rho\Delta_H^R}_{<0, \text{ by high capacity condition}} - w_H\,\rho\,\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\Delta_H^S - w_L(1-\rho)\Delta_L^S < 0 \end{aligned}$$

Thus, $\mathbb{E}\widetilde{W}(\beta)$ is decreasing in β for $\beta \in [\phi(s_L \mid \theta = L), \hat{\beta})$.

Condition 1: $n_0 w_H + l_0 w_L \leq 1$ Given $n_0 w_H + l_0 w_L \leq 1$, the highest expected welfare for $\beta \in [\beta_1, \phi(s_L \mid \theta = L)]$ is achieved at $\phi(s_L \mid \theta = L)$, which is given by:

$$\begin{aligned} \mathbb{E}\widetilde{W}(\phi(s_L \mid \theta = L)) = &\left[-(1-\rho)\Delta_L^R\phi(s_H \mid \theta = L) + \rho\Delta_H^R\phi(s_H|\theta=H)\right] \\ &+ w_H\rho\Delta_H^S\phi(s_H|\theta=H) + w_L(1-\rho)\left(\Delta_L^S\phi(s_H \mid \theta = L) - C\right) \end{aligned}$$

The expected welfare at $\hat{\beta}$ is given by:

$$\mathbb{E}\widetilde{W}(\hat{\beta}) = \rho\Delta_H^R\left(\frac{\phi(s_H \mid \theta = H)}{\phi(s_H \mid \theta = L)} - 1\right)\left(\frac{C}{\Delta_L^S - C}\right) + w_H\rho\Delta_H^S\frac{\phi(s_H \mid \theta = H)}{\phi(s_H \mid \theta = L)}\left(\frac{C}{\Delta_L^S}\right)$$

The difference in expected welfare is:

$$\mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\phi(s_L \mid \theta = L)) = \left[(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^R\right][m_1(C) - n_1(C)w_H - l_1(C)w_L]$$
$$\propto m_1(C) - n_1(C)w_H - l_1(C)w_L$$

where

$$m_1(C) := \frac{\mathbb{E}U^R\left(\hat{\beta}\right) - \mathbb{E}U^R\left(\phi(s_L \mid \theta = L)\right)}{(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R}$$
$$= \frac{\rho\Delta_H^R\left(\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - 1\right)\left(\frac{C}{\Delta_L^S - C}\right) + (1-\rho)\Delta_L^R\phi(s_H \mid \theta = L) - \rho\Delta_H^R\phi(s_H|\theta = H)}{(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R}$$

$$n_1(C) := \frac{\rho\frac{\Delta_H^S}{\Delta_L^S}\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\left(\Delta_L^S\phi(s_H \mid \theta = L) - C\right)}{(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R},\ l_1(C) := \frac{(1-\rho)\left(\Delta_L^S\phi(s_H \mid \theta = L) - C\right)}{(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R}$$

By the definition of $\hat{C}$ that it is the value of C that make the receiver indifferent between $\hat{\beta}$ and $\phi(s_L \mid \theta = L)$, the zero point of $m_1(C)$ is the $\hat{C}$. As $m_1(C)$ is increasing in C, we have $m_1(C) \geq m_1(\hat{C}) = 0$ for $C \in [\hat{C}, \Delta_L^S\phi(s_H \mid \theta = L))$.

By $\hat{\beta} := 1 - C/\Delta_L^S > \phi(s_L \mid \theta = L)$, we have $C < \Delta_L^S\phi(s_H \mid \theta = L)$. Thus, $n_1(C) > 0$ and $l_1(C) > 0$.

Condition 2: $n_0w_H + l_0w_L > 1$ Given $n_0w_H + l_0w_L > 1$, the highest expected welfare for $\beta \in [\beta_1, \phi(s_L \mid \theta = L)]$ is achieved at β_1, which is given by:

$$\mathbb{E}\widetilde{W}(\beta_1) = \mathbb{E}\widetilde{W}(\phi(s_L \mid \theta = L))$$
$$+ (n_0w_H + l_0w_L - 1)\left[(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^R\right](\phi(s_L \mid \theta = L) - \beta_1)$$

The difference in expected welfare of β_1 and $\hat{\beta}$ is:

$$\mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\beta_1) = \mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\phi(s_L \mid \theta = L))$$
$$- (n_0w_H + l_0w_L - 1)\left[(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^R\right](\phi(s_L \mid \theta = L) - \beta_1)$$
$$= \left[(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^R\right][m_2(C) - w_Hn_2(C) - w_Ll_2(C)]$$

where

$$m_2(C) = m_1(C) + (\phi(s_L \mid \theta = L) - \beta_1)$$
$$n_2(C) = n_1(C) - n_0(\phi(s_L \mid \theta = L) - \beta_1)$$
$$l_2(C) = l_1(C) - l_0(\phi(s_L \mid \theta = L) - \beta_1)$$

Then, $m_2(C) - w_H n_2(C) - w_L l_2(C) = m_1(C) - n_1(C)w_H - l_1(C)w_L - (n_0 w_H + l_0 w_L - 1)(\phi(s_L \mid \theta = L) - \beta_1)$. Then,

$$\mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\beta_1) \propto m_1(C) - n_1(C)w_H - l_1(C)w_L - (n_0 w_H + l_0 w_L - 1)(\phi(s_L \mid \theta = L) - \beta_1)$$

Optimal β: Based on the derivation of Condition 1 and 2,

$$\mathbb{E}\widetilde{W}(\hat{\beta}) - \max_{\beta \in [\beta_1, \phi(s_L|\theta=L)]} \mathbb{E}\widetilde{W}(\beta) \propto M(C; w_H, w_L)$$

where

$$M(C; w_H, w_L) := m_1(C) - n_1(C)w_H - l_1(C)w_L - (n_0 w_H + l_0 w_L - 1)^+(\phi(s_L \mid \theta = L) - \beta_1),$$

and it is increasing in C and decreasing in w_H and w_L, with $M(\hat{C}; w_H, w_L) = -n_1(\hat{C})w_H - l_1(\hat{C})w_L - (n_0 w_H + l_0 w_L - 1)^+(\phi(s_L \mid \theta = L) - \beta_1) < 0$ for any w_H and w_L.

Because $m_1(C) - n_1(C)w_H - l_1(C)w_L$ is increasing in C and

$$m_1(C) - n_1(C)w_H - l_1(C)w_L\Big|_{C=\Delta_L^S \phi(s_H|\theta=L)} = m_1(\Delta_L^S \phi(s_H \mid \theta = L)) = \phi(s_H \mid \theta = L) > 0$$

The $M(C; w_H, w_L) < 0$ holds for any $C \in [\hat{C}, \Delta_L^S \phi(s_H \mid \theta = L))$ if and only if

$$n_0 w_H + l_0 w_L \geq \frac{m_1(\Delta_L^S \phi(s_H \mid \theta = L))}{\phi(s_L \mid \theta = L) - \beta_1} + 1 = \frac{1 - \beta_1}{\phi(s_L \mid \theta = L) - \beta_1} > 1$$

The set of the optimal true-positive rate β^* is given as follows:

1. If $n_0 w_H + l_0 w_L < (1 - \beta_1)/[\phi(s_L \mid \theta = L) - \beta_1]$, the zero point of $M(C; w_H, w_L)$ in the interval $[\hat{C}, \Delta_L^S \phi(s_H \mid \theta = L))$ is denoted as $C_1^*(w_H, w_L)$. The set of the optimal true-positive rate β^* is given by

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C \in [\hat{C}, C_1^*(w_H, w_L)) \\ \mathcal{B}(w_H, w_L) \cup \{\hat{\beta}\} & \text{if } C = C_1^*(w_H, w_L) \\ \{\hat{\beta}\} & \text{if } C \in (C_1^*(w_H, w_L), \Delta_L^S \phi(s_H \mid \theta = L)) \end{cases}$$

By implicit function theorem,

$$\frac{\partial C_1^*(w_H, w_L)}{\partial w_H} = -\frac{\left.\frac{\partial M(C;w_H,w_L)}{\partial w_H}\right|_{C=C_1^*(w_H,w_L)}}{\left.\frac{\partial M(C;w_H,w_L)}{\partial C}\right|_{C=C_1^*(w_H,w_L)}} > 0,$$

$$\frac{\partial C_1^*(w_H, w_L)}{\partial w_L} = -\frac{\left.\frac{\partial M(C;w_H,w_L)}{\partial w_L}\right|_{C=C_1^*(w_H,w_L)}}{\left.\frac{\partial M(C;w_H,w_L)}{\partial C}\right|_{C=C_1^*(w_H,w_L)}} > 0$$

2. If $n_0 w_H + l_0 w_L \geq (1-\beta_1)/[\phi(s_L \mid \theta = L) - \beta_1]$, the set of the optimal true-positive rate β^* is $\{\beta_1\}$.

A.10.3 High lying cost regime: $C \in [\Delta_L^S \phi(s_H \mid \theta = L), \Delta_L^S(1-\beta_1)]$

This corresponds to $\hat{\beta} \leq \phi(s_L \mid \theta = L)$. The optimal β lies in the interval $[\beta_1, \phi(s_L \mid \theta = L)]$. The expected welfare is given by

$$\mathbb{E}\widetilde{W}(\beta) = \begin{cases} \rho(1 - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\beta)(\Delta_H'^S + \Delta_H^R) + (1-\rho)\left[(1-\beta)(\Delta_L'^S - \Delta_L^R) - C'\right], & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \rho\left(1 - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\right)(\Delta_H'^S + \Delta_H^R) + \rho C' \frac{\Delta_H'^S}{\Delta_L'^S}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}, & \beta \in [\hat{\beta}, \phi(s_L \mid \theta = L)] \end{cases}$$

The $\mathbb{E}\widetilde{W}(\beta)$ is constant for $\beta \in [\hat{\beta}, \phi(s_L \mid \theta = L)]$. For $\beta \in \left[\beta_1, \hat{\beta}\right)$, the form of $\mathbb{E}\widetilde{W}(\beta)$ is the same as it within the low lying cost regime. Thus, the $\partial \mathbb{E}\widetilde{W}(\beta)/\partial\beta$ is proportional to $1 - (n_0 w_H + l_0 w_L)$. Thus, if $n_0 w_H + l_0 w_L \leq 1$, all $\beta \in \left[\beta_1, \hat{\beta}\right)$ is dominated by any $\beta \in [\hat{\beta}, \phi(s_L \mid \theta = L)]$.

If $n_0 w_H + l_0 w_L > 1$, we need to compare $\mathbb{E}\widetilde{W}(\beta_1)$ and $\mathbb{E}\widetilde{W}(\hat{\beta})$, where

$$\begin{aligned} \mathbb{E}\widetilde{W}(\beta_1) &= \rho\Delta_H^R(1 - \beta_1\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}) - (1-\beta_1)(1-\rho)\Delta_L^R \\ &\quad + w_L(1-\rho)[(1-\beta_1)\Delta_L^S - C] + w_H\rho\Delta_H^S(1 - \beta_1\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}) \\ \mathbb{E}\widetilde{W}(\hat{\beta}) &= \rho\left(1 - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\right)(w_H\Delta_H^S + \Delta_H^R) + w_H\rho C\frac{\Delta_H^S}{\Delta_L^S}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} \end{aligned}$$

The difference in expected welfare between $\hat{\beta}$ and β_1 is:

$$\begin{aligned} \mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\beta_1) &= \underbrace{(\beta_1 - 1)\left[-(1-\rho)\Delta_L^R + \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^R\right]}_{>0} \\ &\quad + w_H\rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^S(\beta_1 - \hat{\beta}) \\ &\quad + w_L(1-\rho)\Delta_L^S(\beta_1 - \hat{\beta}) \end{aligned}$$

Then, $\mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\beta_1) \leq 0$ can be given by

$$m_3 w_H + n_3 w_L \geq 1$$

where

$$m_3 := \frac{\rho \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} \Delta_H^S (\hat{\beta} - \beta_1)}{(\beta_1 - 1)\left[-(1-\rho)\Delta_L^R + \rho \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} \Delta_H^R\right]} = \frac{\hat{\beta} - \beta_1}{1 - \beta_1} n_0 > 0,$$

$$n_3 := \frac{(1-\rho)\Delta_L^S (\hat{\beta} - \beta_1)}{(\beta_1 - 1)\left[-(1-\rho)\Delta_L^R + \rho \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} \Delta_H^R\right]} = \frac{\hat{\beta} - \beta_1}{1 - \beta_1} l_0 > 0$$

Thus, $m_3 w_H + n_3 w_L \geq 1$ is equivalent to

$$n_0 w_H + l_0 w_L \geq \frac{1 - \beta_1}{\hat{\beta} - \beta_1}$$

and which is also equivalent to

$$C \leq (1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right) \Delta_L^S$$

Note that $\hat{\beta} \leq \phi(s_L \mid \theta = L)$, we have $n_0 w_H + l_0 w_L \leq (1 - \beta_1)/(\hat{\beta} - \beta_1)$ holds for all $C \in [\Delta_L^S \phi(s_H \mid \theta = L), \Delta_L^S (1 - \beta_1)]$ if and only if $n_0 w_H + l_0 w_L \leq (1 - \beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$.

The set of the optimal true-positive rate β^* is given as follows:

1. If $n_0 w_H + l_0 w_L < (1 - \beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$, the set of the optimal true-positive rate β^* is $[\hat{\beta}, \phi(s_L \mid \theta = L)]$.

2. If $n_0 w_H + l_0 w_L \geq (1 - \beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$, the set of the optimal true-positive rate β^* is given by

$$\begin{cases} \{\beta_1\} & \text{if } C \in \left[\Delta_L^S \phi(s_H \mid \theta = L), (1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right) \Delta_L^S\right) \\ \{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right) \Delta_L^S \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in \left((1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right) \Delta_L^S, \Delta_L^S (1 - \beta_1)\right] \end{cases}$$

A.10.4 Summary

In summary, the set of the optimal true-positive rate β^* is given by

1. **Case 1** $n_0 w_H + l_0 w_L < (1 - \beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$**:** the set of the optimal true-positive rate β^*

is

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < C_1^*(w_H, w_L) \\ \mathcal{B}(w_H, w_L) \cup \{\hat{\beta}\} & \text{if } C = C_1^*(w_H, w_L) \\ \{\hat{\beta}\} & \text{if } C \in (C_1^*(w_H, w_L), \Delta_L^S \phi(s_H \mid \theta = L)) \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in [\Delta_L^S \phi(s_H \mid \theta = L), \Delta_L^S (1 - \beta_1)] \end{cases}$$

$$= \begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < C_1^*(w_H, w_L) \\ \mathcal{B}(w_H, w_L) \cup \{\hat{\beta}\} & \text{if } C = C_1^*(w_H, w_L) \\ [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}] & \text{if } C \in (C_1^*(w_H, w_L), \Delta_L^S (1 - \beta_1)] \end{cases}$$

2. **Case 2** $n_0 w_H + l_0 w_L \geq (1 - \beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$**:** the set of the optimal true-positive rate β^* is

$$\begin{cases} \{\beta_1\} & \text{if } C < (1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\ \{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in \left((1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S, \Delta_L^S (1 - \beta_1)\right] \end{cases}$$

where

$$\mathcal{B}(w_H, w_L) := \begin{cases} \{\phi(s_L \mid \theta = L)\} & \text{if } n_0 w_H + l_0 w_L < 1 \\ [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } n_0 w_H + l_0 w_L = 1 \\ \{\beta_1\} & \text{if } n_0 w_H + l_0 w_L > 1 \end{cases}$$

By the definition of $C_1^*(w_H, w_L)$, we can define a continuous function $\tilde{C}(w_H, w_L)$ as follows:

$$\tilde{C}(w_H, w_L) := \begin{cases} C_1^*(w_H, w_L), & \text{if } n_0 w_H + l_0 w_L < \frac{1-\beta_1}{\phi(s_L|\theta=L)-\beta_1} \\ (1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S, & \text{if } n_0 w_H + l_0 w_L \geq \frac{1-\beta_1}{\phi(s_L|\theta=L)-\beta_1} \end{cases},$$

which is increasing in w_H and w_L. Then, the set of the optimal true-positive rate β^* can be written as

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < \tilde{C}(w_H, w_L) \\ \mathcal{B}(w_H, w_L) \cup [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}] & \text{if } C = \tilde{C}(w_H, w_L) \\ [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}] & \text{if } C \in (\tilde{C}(w_H, w_L), \Delta_L^S (1 - \beta_1)] \end{cases}$$

A.11 Proof of Proposition 6

According to Lemma 5, we must have the constraints on the detector with $\{\beta, \alpha^*(\beta, \phi)\}$, where $\beta \leq \phi(s_L \mid \theta = L)$ and $\alpha^*(\beta, \phi) = (\phi(s_L|\theta = H)/\phi(s_L|\theta = L))\beta$. That is, compared to the original model, the only difference in the extension is the constraint on the range of β.

Case 1: Low-capacity classifier According to the proof of Proposition 3 in Appendix A.8, if the classifier has a low capacity, we have $\mathbb{E}U^R(\beta) = 0$ when $\beta < \hat{\beta}$. When $\beta \geq \hat{\beta}$, we have $\mathbb{E}U^R(\beta) = [1 -$

$\alpha^*(\beta;\phi)/\beta]\rho\Delta_H^R$, which is constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta=L)\}]$.

Therefore, the set of optimal true-positive rates that maximize the receiver's expected payoff is:

- $[0, \phi(s_L|\theta=L)]$ if $\phi(s_L|\theta=L) < \hat{\beta}$ (i.e., $C < \Delta_L^S \phi(s_H \mid \theta=L)$), which gives zero payoff to the receiver
- $[\hat{\beta}, \phi(s_L|\theta=L)]$ if $\phi(s_L|\theta=L) \geq \hat{\beta}$ (i.e., $C \geq \Delta_L^S \phi(s_H \mid \theta=L)$), which gives a constant payoff $[1-\phi(s_L|\theta=H)/\phi(s_L|\theta=L)]\rho\Delta_H^R$ to the receiver

By the proof of Proposition 4 in Appendix A.9, if the classifier has a low capacity, the payoff of the low-type sender is $\mathbb{E}U_L^S(\beta) = 0$ and the payoff of the high-type sender is

$$\mathbb{E}U_H^S(\beta) = \begin{cases} C\frac{\Delta_H^S}{\Delta_L^S}\frac{1-\alpha^*(\beta;\phi)}{1-\beta}, & \beta \in \left[0, \hat{\beta}\right) \\ \Delta_H^S - (\Delta_L^S - C)\frac{\Delta_H^S}{\Delta_L^S}\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

Thus, the set of optimal true-positive rates that maximize the high-type sender's payoff is:

- $\{\phi(s_L|\theta=L)\}$ if $\phi(s_L|\theta=L) < \hat{\beta}$ (i.e., $C < \Delta_L^S \phi(s_H \mid \theta=L)$)
- $[\hat{\beta}, \phi(s_L|\theta=L)]$ if $\phi(s_L|\theta=L) \geq \hat{\beta}$ (i.e., $C \geq \Delta_L^S \phi(s_H \mid \theta=L)$)

For the weighted sum of the receiver's and the sender's payoffs, $\mathbb{E}\widetilde{W}(\beta) = \mathbb{E}U^R(\beta) + w_L(1-\rho)\mathbb{E}U_L^S(\beta) + \rho w_H \mathbb{E}U_H^S(\beta)$, the set of optimal true-positive rates is:

- $\{\phi(s_L|\theta=L)\}$ if $\phi(s_L|\theta=L) < \hat{\beta}$ (i.e., $C < \Delta_L^S \phi(s_H \mid \theta=L)$)
- $[\hat{\beta}, \phi(s_L|\theta=L)]$ if $\phi(s_L|\theta=L) \geq \hat{\beta}$ (i.e., $C \geq \Delta_L^S \phi(s_H \mid \theta=L)$)

Case 2: High-capacity classifier: $\phi(s_H|\theta=H)/\phi(s_H|\theta=L) \geq [(1-\rho)\Delta_L^R]/(\rho\Delta_H^R)$ **and** $C \leq (1-\beta_1)\Delta_L^S$ According to the proof of Proposition 3 in Appendix A.8, if the classifier has a high capacity, the receiver's payoff is

$$\mathbb{E}U^R(\beta) = \begin{cases} 0, & \beta \in [0, \beta_1) \\ (1-\alpha^*(\beta;\phi))\rho\Delta_H^R - (1-\beta)(1-\rho)\Delta_L^R, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \left(1 - \frac{\alpha^*(\beta;\phi)}{\beta}\right)\rho\Delta_H^R, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

By Lemma 12 that $\phi(s_L|\theta=L) \geq \beta_1$, the set of optimal true-positive rates that maximize the receiver's payoff is

$$\begin{cases} [0, \phi(s_L|\theta=L)] & \text{if } \phi(s_L|\theta=L) < \hat{\beta} \text{ and } \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} = \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} \\ \{\phi(s_L|\theta=L)\} & \text{if } \phi(s_L|\theta=L) < \hat{\beta} \text{ and } \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} > \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} \\ [\hat{\beta}, \phi(s_L|\theta=L)] & \text{if } \phi(s_L|\theta=L) \geq \hat{\beta} \end{cases}$$

By the proof of Proposition 4 in Appendix A.9, if the classifier has a high capacity, the low-type sender's payoff is

$$\mathbb{E}U_L^S(\beta) = \begin{cases} 0, & \beta \in [0, \beta_1) \\ (1-\beta)\Delta_L^S - C, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ 0, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

and the high-type sender's payoff is

$$\mathbb{E}U_H^S(\beta) = \begin{cases} C\frac{\Delta_H^S}{\Delta_L^S}\frac{1-\alpha^*(\beta;\phi)}{1-\beta} & \beta \in [0, \beta_1) \\ (1-\alpha^*(\beta;\phi))\Delta_H^S, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \Delta_H^S - (\Delta_L^S - C)\frac{\Delta_H^S}{\Delta_L^S}\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

which is maximized at $\beta = \beta_1$ by Proposition 4. Since $\phi(s_L|\theta = L) \geq \beta_1$ by Lemma 12, $\mathbb{E}U_H^S(\beta)$ is also maximized at $\beta = \beta_1$ in this case.

For the weighted sum of payoffs $\mathbb{E}\widetilde{W}(\beta) = \mathbb{E}U^R(\beta) + w_L(1-\rho)\mathbb{E}U_L^S(\beta) + \rho w_H \mathbb{E}U_H^S(\beta)$, according to the proof of Proposition 5 in Appendix A.10, we have the following results:

For $C < \Delta_L^S \phi(s_H \mid \theta = L)$ (i.e., $\phi(s_L|\theta = L) < \hat{\beta}$), the set of optimal true-positive rates is given by

$$\mathcal{B}(w_H, w_L) := \begin{cases} \{\phi(s_L \mid \theta = L)\} & \text{if } n_0 w_H + l_0 w_L < 1 \\ \{\beta_1\} & \text{if } n_0 w_H + l_0 w_L > 1 \\ [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } n_0 w_H + l_0 w_L = 1 \end{cases}$$

For $C \geq \Delta_L^S \phi(s_H \mid \theta = L)$ (i.e., $\phi(s_L|\theta = L) \geq \hat{\beta}$), the analysis is more complex:

1. If $n_0 w_H + l_0 w_L < (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$, the set of optimal true-positive rates is $[\hat{\beta}, \phi(s_L \mid \theta = L)]$.

2. If $n_0 w_H + l_0 w_L \geq (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$, the set of optimal true-positive rates is

$$\begin{cases} \{\beta_1\} & \text{if } C \in \left[\Delta_L^S \phi(s_H \mid \theta = L), (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S\right) \\ \{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in \left((1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S, \Delta_L^S(1-\beta_1)\right] \end{cases}$$

Summary: The set of optimal true-positive rates β^* is characterized as follows:

1. **Case 1:** $n_0 w_H + l_0 w_L < (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < \Delta_L^S \phi(s_H \mid \theta = L) \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in [\Delta_L^S \phi(s_H \mid \theta = L), \Delta_L^S(1-\beta_1)] \end{cases}$$

2. **Case 2:** $n_0 w_H + l_0 w_L \geq (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$

$$\begin{cases} \{\beta_1\} & \text{if } C < (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\ \{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in \left((1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S, \Delta_L^S(1-\beta_1)\right] \end{cases}$$

Combining the results from the two cases, the set of optimal true-positive rates β^* is characterized as follows:

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < \max\left\{\Delta_L^S \phi(s_H \mid \theta = L), (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S\right\} \\ \{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S > \Delta_L^S \phi(s_H \mid \theta = L) \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{Otherwise} \end{cases}$$

A.12 Proof of Proposition 7

Equilibrium First, if $f > 1 - u_H/P$, high-quality sellers do not enter the market because the maximum profit from entering, $(1-f)P$, is lower than the reservation utility, u_H. In this case, low-quality sellers also do not enter the market because they cannot induce any transactions when consumers know there are no high-quality sellers. The market breaks down.

Second, if $f \in (1-(C+u_L)/P, 1-u_H/P]$, low-quality sellers do not enter the market because the maximum profit from entering, $-C+(1-f)P$, is less than the reservation utility, u_L, and only high-quality sellers may enter the market. Thus, the optimal commission fee above $1-(C+u_L)/P$ is the highest possible f that incentivizes high-quality sellers to enter the market, $1-u_H/P$, which generates a profit of $\Pi = P(1-u_H/P)\rho$.

Third, we consider the case where $f \leq \min\{1-(C+u_L)/P, 1-u_H/P\}$. In such cases, both types of sellers may enter the market. There are four possible (pure strategy) entry decisions:

1. Only low-quality sellers enter the market. This cannot be an equilibrium because consumers will never purchase a product.

2. Only high-quality sellers enter the market. The platform's profit is $\Pi = Pf\rho$.

3. None of the sellers enter the market. This cannot be an equilibrium because high-quality sellers can profitably deviate by entering the market.

4. Both types of sellers enter the market. Upon entry, the game coincides with that in the main model where $\Delta_H^S = \Delta_L^S = (1-f)P$, $\Delta_H^R = 1-P$, and $\Delta_L^R = P - v$. In this case, each high-type seller's sales are $\mathbb{E}U_H^S/(1-f)$ and each low-type seller's sales are $(\mathbb{E}U_L^S + C\sigma^S)/(1-f)$. The platform earns f fractions of the total sales. Thus, its profit is $\Pi = f[\rho\mathbb{E}U_H^S/(1-f) + (1-\rho)(\mathbb{E}U_L^S + C\sigma^S)/(1-f)] = [f/(1-f)][\rho\mathbb{E}U_H^S + (1-\rho)(\mathbb{E}U_L^S + C\sigma^S)]$. Table 4 presents the equilibrium payoffs of the platform and senders for a given commission rate f and detector (β, α).

   It is an equilibrium for both types of sellers to enter the market if and only if the following conditions hold:

β Range \ α Range	$\alpha \leq 1 - \frac{(1-\rho)(P-v)}{\rho(1-P)}(1-\beta)$	$\alpha \in \left(1 - \frac{(1-\rho)(P-v)}{\rho(1-P)}(1-\beta), \beta\right]$
$[1 - C/[(1-f)P], 1)$	$\Pi = \frac{f}{1-f}\rho\left[(1-f)P\left(1-\frac{\alpha}{\beta}\right) + C\frac{1-v}{P-v}\frac{\alpha}{\beta}\right]$, $\mathbb{E}U_L^S = 0$, $\mathbb{E}U_H^S = (1-f)P\left(1-\frac{\alpha}{\beta}\right) + C\frac{\alpha}{\beta}$	
$(0, 1 - C/[(1-f)P])$	$\Pi = f[\rho(1-\alpha) + (1-\rho)(1-\beta)]P$, $\mathbb{E}U_L^S = (1-\beta)(1-f)P - C$, $\mathbb{E}U_H^S = (1-\alpha)(1-f)P$	$\Pi = \frac{f}{1-f}\rho C\frac{1-v}{P-v}\frac{1-\alpha}{1-\beta}$, $\mathbb{E}U_L^S = 0$, $\mathbb{E}U_H^S = C\frac{1-\alpha}{1-\beta}$

Table 4: Equilibrium Payoffs for a Given Commission Rate and Detector

(a) A low-quality seller's equilibrium payoff exceeds his reservation utility, $\mathbb{E}U_L^S \geq u_L$.

(b) A high-quality seller's equilibrium payoff exceeds his reservation utility, $\mathbb{E}U_H^S \geq u_H$.

According to Table 4, the above conditions are equivalent to $\alpha \leq 1-(1-\beta)(1-\rho)(P-v)/[\rho(1-P)]$, $\alpha \leq 1 - u_H/[(1-f)P]$, and $\beta \leq 1 - (C+u_L)/[(1-f)P]$. Following Section 5.1, we analyze the case of $\lambda_H = 0$, where $\beta \leq \phi(s_L|\theta = L)$ and $\alpha = \alpha^*(\beta, \phi) = \beta\phi(s_L|\theta = H)/\phi(s_L|\theta = L)$. The designer's optimization problem is:

$$\begin{aligned}
&\max_{f,\beta,\alpha} f[\rho(1-\alpha) + (1-\rho)(1-\beta)]P \qquad (3)\\
&\quad \text{s.t. } \beta \leq \phi(s_L|\theta = L), \alpha = \alpha^*(\beta, \phi),\\
&\qquad \alpha \leq 1 - \frac{(1-\rho)(P-v)}{\rho(1-P)}(1-\beta), \alpha \leq 1 - \frac{u_H}{(1-f)P}, \text{ and } \beta \in \left(0, 1 - \frac{C+u_L}{(1-f)P}\right]\\
\Leftrightarrow\ &\max_{f,\beta} f\left[\rho\left(1 - \frac{\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)}\beta\right) + (1-\rho)(1-\beta)\right]P\\
&\quad \text{s.t. } \eta \leq \beta \leq \phi(s_L|\theta = L)\\
&\qquad \beta \leq \frac{\phi(s_L|\theta = L)}{\phi(s_L|\theta = H)}\left[1 - \frac{u_H}{(1-f)P}\right], \text{ and } \beta \leq 1 - \frac{C+u_L}{(1-f)P},\\
&\quad \text{where } \eta := \left[\frac{(1-\rho)(P-v)}{\rho(1-P)} - 1\right] / \left[\frac{(1-\rho)(P-v)}{\rho(1-P)} - \frac{\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)}\right]
\end{aligned}$$

Notice that the upper bounds on β decrease in f. So, $f = 0$ corresponds to the most relaxed constraints on β. The constraints of the designer's optimization problem can be satisfied if and only if there exists a feasible β when $f = 0$, which is equivalent to

$$\eta \leq \phi(s_L|\theta = L) \left(\Leftrightarrow \frac{\phi(s_H|\theta = H)}{\phi(s_H|\theta = L)} \geq \frac{(1-\rho)(P-v)}{\rho(1-P)}\right) \qquad (4)$$

$$\text{and } \eta \leq \min\left\{\frac{\phi(s_L|\theta = L)}{\phi(s_L|\theta = H)}(1 - \frac{u_H}{P}), 1 - \frac{C+u_L}{P}\right\} \qquad (5)$$

Condition (5) is more likely to be satisfied when the classifier has higher capacity (i.e., when $\phi(s_L|\theta =$

$L)/\phi(s_L|\theta = H)$ is larger) because η decreases in $\phi(s_L|\theta = L)/\phi(s_L|\theta = H)$ whereas the right-hand side of the inequality increases in $\phi(s_L|\theta = L)/\phi(s_L|\theta = H)$. In the limiting case where $\phi(s_L|\theta = L)/\phi(s_L|\theta = H) \to +\infty$, the condition reduces to $(C + u_L)/[(1 - \bar{\eta})P] \leq 1$, where $\bar{\eta} := 1 - \rho(1 - P)/[(1 - \rho)(P - v)]$. When $(C + u_L)/[(1 - \bar{\eta})P] < 1$, there exists a threshold ϕ_1^* such that (5) holds if and only if $\phi(s_L|\theta = L)/\phi(s_L|\theta = H) \geq \phi_1^*$.

When both conditions (4) and (5) hold, the optimal detector that induces both types of sellers to enter the market is

$$\beta^* = \eta, \alpha^* = \alpha^*(\eta, \phi) = \frac{\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)}\eta, \text{ and } f^* = 1 - \max\left\{\frac{u_H}{1 - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\eta}, \frac{C + u_L}{1 - \eta}\right\}\frac{1}{P}.$$

The optimal true-positive rate $\beta^* = \eta$ decreases in $\phi(s_L|\theta = L)/\phi(s_L|\theta = H)$. The corresponding platform's profit is $\Pi_2^* = f^* \cdot P \cdot \{\rho[1 - \eta\phi(s_L|\theta = H)/\phi(s_L|\theta = L)] + (1 - \rho)(1 - \eta)\}$, which is strictly increasing in $\phi(s_L|\theta = L)/\phi(s_L|\theta = H)$.

Optimal Commission Rate and Detector We just need to compare the profit under the optimal commission rate and detector that induces only high-quality sellers to enter the market with the profit under the optimal commission rate and detector that induces both types of sellers to enter the market.

The platform obtains the highest profit when only high-quality sellers enter the market, $\Pi_1^* = P(1 - u_H/P)\rho$, by setting $f = 1 - u_H/P$. In this case, one can verify that conditions (4) and (5) cannot be simultaneously satisfied. Thus, only high-quality sellers enter the market given any detector. The choice of the detector does not affect the platform's profits.

When $\phi(s_L|\theta = L)/\phi(s_L|\theta = H) \to +\infty$, the platform's profit becomes

$$\left[1 - \max\left\{\frac{u_H}{P}, \frac{C + u_L}{(1 - \bar{\eta})P}\right\}\right][\rho + (1 - \rho)(1 - \bar{\eta})]\,P$$

This exceeds Π_1^* if and only if $(C + u_L)/[(1 - \bar{\eta})P] \leq 1 - (1 - u_H/P)\rho/[\rho + (1 - \rho)(1 - \bar{\eta})] \Leftrightarrow C < \rho(1 - P)P[1 - (1 - u_H/P)(P - v)/(1 - v)]/[(1 - \rho)(P - v)] - u_L$. Under the above condition, there exists ϕ_2^* such that Π_2^* exceeds Π_1^* if and only if $\phi(s_L|\theta = L)/\phi(s_L|\theta = H) > \phi_2^*$.

We conclude the proof by defining $\bar{\phi} := \max\{\phi_1^*, \phi_2^*\}$.